Available online at www.sciencedirect.com
ScienceDirect
Journal of Econometrics 142 (2008) 134–161
JOURNAL OF Econometrics
www.elsevier.com/locate/jeconom

Generalized empirical likelihood tests in time series models with potential identification failure

Patrik Guggenberger[a,*], Richard J. Smith[b,c]

[a]*Department of Economics, U.C.L.A., Bunche Hall 8385, Box 951477, Los Angeles, CA 90095-1477, USA*
[b]*Cemmap, U.C.L and I.F.S., UK*
[c]*Faculty of Economics, University of Cambridge, UK*

Received 27 April 2005; accepted 29 March 2007
Available online 2 June 2007

Abstract

We introduce test statistics based on generalized empirical likelihood methods that can be used to test simple hypotheses involving the unknown parameter vector in moment condition time series models. The test statistics generalize those in Guggenberger and Smith [2005. Generalized empirical likelihood estimators and tests under partial, weak and strong identification. Econometric Theory 21 (4), 667–709] from the i.i.d. to the time series context and are alternatives to those in Kleibergen [2005a. Testing parameters in GMM without assuming that they are identified. Econometrica 73 (4), 1103–1123] and Otsu [2006. Generalized empirical likelihood inference for nonlinear and time series models under weak identification. Econometric Theory 22 (3), 513–527]. The main feature of these tests is that their empirical null rejection probabilities are not affected much by the strength or weakness of identification. More precisely, we show that the statistics are asymptotically distributed as chi-square under both classical asymptotic theory and weak instrument asymptotics of Stock and Wright [2000. GMM with weak identification. Econometrica 68 (5), 1055–1096]. We also introduce a modification to Otsu's (2006) statistic that is computationally more attractive. A Monte Carlo study reveals that the finite-sample performance of the suggested tests is very competitive.

JEL classification: C12; C31

Keywords: Generalized empirical likelihood; Nonlinear moment conditions; Similar tests; Size distortion; Weak identification

1. Introduction

There has recently been a lot of interest in robust inference in weakly identified models.[1] This paper adds to this literature by introducing two types of test statistics that can be used to test simple hypotheses involving the

*Corresponding author. Tel.: +1 310 825 0849; fax: +1 310 825 9528.
E-mail address: guggenbe@econ.ucla.edu (P. Guggenberger).

[1]See inter alia Dufour (1997), Staiger and Stock (1997), Stock and Wright (2000), Kleibergen (2002, 2005a), Caner (2003), Moreira (2003), Andrews and Marmer (2004), Chao and Swanson (2005), Dufour and Taamouti (2005), Guggenberger and Smith (2005), Moreira et al. (2005a,b), Andrews et al. (2006), and Otsu (2006). For a recent discussion of that literature, see Dufour (2003). For a general theory of the size of tests in situations where the asymptotic distribution of the test statistic is discontinuous in nuisance parameters, see Andrews and Guggenberger (2005a–c).

0304-4076/$ - see front matter
doi:10.1016/j.jeconom.2007.03.003

unknown parameter vector in *nonlinear* moment condition *time series* models. The main feature of these statistics is that they lead to tests whose empirical rejection probabilities (ERPs) under the null hypothesis do not depend much on the strength or weakness of identification of the model. More precisely, we show that the statistics are asymptotically distributed as chi-square under both classical and the weak instrument asymptotic theory of Stock and Wright (2000). This is in contrast to many of the classical test statistics, like, for example, Wald statistics, that have a chi-square under the former but a nonstandard asymptotic distribution under the latter theory.

The first test statistic is given as the renormalized criterion function of the generalized empirical likelihood (GEL) estimator, see Smith (1997, 2001) and Newey and Smith (2004), and the second one as a quadratic form in the first-order condition (FOC) of the GEL estimator; both statistics are evaluated at the hypothesized parameter vector. The statistics generalize those in Guggenberger and Smith (2005) (GS henceforth) from the i.i.d. and martingale difference sequence (m.d.s.) setup to the time series case. One advantage of the second statistic over the first one is that the degrees of freedom parameter of its asymptotic chi-square distribution equals p, the dimension of the unknown parameter vector, while for the first statistic the degrees of freedom parameter equals k, the number of moment conditions. This negatively affects power properties of tests based on the first statistic in over-identified situations. To adapt the statistics to the time series context, we work with smoothed counterparts of the moment indicator functions based on a kernel function $k(\cdot)$ and a bandwidth parameter S_n, an approach which was originally used in Kitamura and Stutzer (1997) and Smith (1997, 2001). This method for the construction of test statistics in the weakly identified framework was suggested by Guggenberger (2003, Introduction of the first chapter). See also Otsu (2006). To clarify the need for smoothing, we also derive the non-pivotal limit distributions of the unsmoothed statistics in GS in the weak identification time series context considered here.

While most of the papers on robust testing with weak identification are written for the linear i.i.d. instrumental variables (IV) model, there are two closely related procedures for robust inference in nonlinear time series models available in the literature. Firstly, Kleibergen (2005a) introduces a test statistic that is given as a quadratic form in the FOC of the generalized method of moments (GMM, Hansen, 1982) continuous updating estimator (CUE). The statistic includes consistent estimators for the long-run covariance matrix of the sums of the renormalized moment indicators and derivatives thereof. Kleibergen (2005a) suggests the use of heteroskedasticity and autocorrelation consistent (HAC) estimators, see Andrews (1991). Secondly, Otsu's (2006) procedure is based on the criterion function of the GEL estimator. An asymptotic chi-square null distribution with p degrees of freedom is obtained by evaluating the GEL criterion function at transformed moment indicators of dimension p rather than at the original moment indicators that are k-dimensional. In Section 2.4 below we give a detailed comparison of the various approaches. There we also introduce modifications to Otsu's (2006) statistic that are computationally more attractive and two hybrid statistics that can be viewed as compromises between our GEL-type and Kleibergen's (2005a) GMM-type procedures.

Besides technicalities, the main assumptions needed to establish the asymptotic chi-square null distribution of the new test statistics introduced in this paper are that (1) an appropriate HAC estimator of the long-run covariance matrix of the sums of the moment indicators is consistent and that (2) a central limit theorem (CLT) holds for the moment indicators and derivatives thereof with respect to the weakly identified parameters. These assumptions are very similar to the ones used in Kleibergen (2005a). They are stated and discussed in the Appendix.

The tests in this paper are first introduced for simple hypotheses on the *full* parameter vector. They are then generalized to sub-vector tests under the assumption that the parameters not under test are strongly identified, see e.g. Kleibergen (2004, 2005a), GS, and Otsu (2006). The idea is to replace the parameters not under test by consistently estimated counterparts in the test statistics.

To investigate the finite-sample performance of the new tests, we compare them to those in Kleibergen (2005a) and Otsu (2006) in a comprehensive Monte Carlo study that focuses on a time series linear model with AR(1) or MA(1) variables. We find that both in terms of size and power the new tests compare very favorably to the alternative procedures. Even though the tests are first-order equivalent, there can be huge power differences between Kleibergen's (2005a), Otsu's (2006), and the tests in this paper.

To implement the tests here and those in Kleibergen (2005a) and Otsu (2006) a bandwidth S_n has to be chosen. Andrews (1991) and Newey and West (1994) provide theory of how to choose the bandwidth, if the goal is to minimize the mean-squared error of a (HAC) covariance matrix estimator. However, in the testing context here, we are really interested in size and power properties of the tests and it is unclear how to develop a theory of bandwidth choice. One could still follow the procedures in Andrews (1991) or Newey and West (1994) but very likely this would not lead to any optimality result. The bandwidth choice is an important problem that is beyond the scope of this paper. Future research has to tackle this challenging question.

The remainder of the paper is organized as follows. In Section 2, the model and the full- and sub-vector test statistics are introduced and their asymptotic theory is discussed. The tests are compared to Kleibergen's (2005a) and Otsu's (2006) approaches. Section 3 contains the Monte Carlo study. All technical assumptions and proofs are relegated to the Appendix.

The symbols "$\to_d$" and "$\to_p$" denote convergence in distribution and convergence in probability, respectively. Convergence "almost surely" is written as "a.s." and "with probability approaching 1" is replaced by "w.p.a.1". The space $C^i(S)$ contains all functions that are i-times continuously differentiable on the set S. Furthermore, $vec(M)$ stands for the column vectorization of the $k \times p$ matrix M, i.e. if $M = (m_1, \ldots, m_p)$ then $vec(M) = (m_1', \ldots, m_p')'$, "$M'$" denotes the transpose matrix of M, $(M)_{i,j}$ the element in the ith row and jth column, "$M > 0$" means that M is positive definite, $\lambda_{\min}(M)$ and $\lambda_{\max}(M)$ are the minimum and maximum eigenvalues of M, respectively, and $\|M\| = \sqrt{\lambda_{\max}(M'M)}$. By I_p we denote the p-dimensional identity matrix.

2. Robust testing

2.1. Model and notation

The paper considers models specified by a finite number of moment restrictions. More precisely, let $\{z_i: i = 1, \ldots, n\}$ be $\mathbb{R}^l$-valued time series data, where $n \in \mathbb{N}$ denotes the sample size. Let $g_n: H \times \Theta \to \mathbb{R}^k$, where $H \subset \mathbb{R}^l$ and $\Theta \subset \mathbb{R}^p$ denotes the parameter space. The model has a true parameter θ_0 for which the moment condition

$$\mathrm{E} g_n(z_i, \theta_0) = 0 \tag{2.1}$$

is satisfied. For $g_n(z_i, \theta)$, usually the shorter $g_i(\theta)$ is used. The function g is allowed to depend on the sample size n to model weak identification, see Assumption ID below. For example, consider the i.i.d. linear IV model given by the structural and reduced form equations $y = Y\theta_0 + u$, $Y = Z\Pi + V$, where $y, u \in \mathbb{R}^n$, $Y, V \in \mathbb{R}^{n\times p}$, $Z \in \mathbb{R}^{n\times k}$, and $\Pi \in \mathbb{R}^{k\times p}$. The matrices Y and Z contain the endogenous variables and instrumental variables, respectively. Denote by $Y_i, V_i, Z_i, \ldots$ $(i = 1, \ldots, n)$ the ith row of the matrix $Y, V, Z, \ldots$ written as a column vector. Assume $\mathrm{E}Z_i u_i = 0$ and $\mathrm{E}Z_i V_i' = 0$. The first condition implies that $\mathrm{E}g_i(\theta_0) = 0$, where for each $i = 1, \ldots, n$, $g_i(\theta) := Z_i(y_i - Y_i'\theta)$. Note that in this example $g_i(\theta)$ depends on n if the reduced form coefficient matrix Π is modeled to depend on n, see Stock and Wright (2000), where $\Pi = \Pi_n = (n^{-1/2}\Pi_A, \Pi_B)$ and Π_A and Π_B are fixed matrices with p_A and p_B columns, $p = p_A + p_B$, and Π_B has full column rank.

Interest focuses on testing a simple hypothesis

$$\mathrm{H}_0:\ \theta_0 = \overline{\theta} \quad \text{versus the alternative} \quad \mathrm{H}_1:\ \theta_0 \neq \overline{\theta}. \tag{2.2}$$

Define the recentered and rescaled sample average

$$\Psi_n(\theta) := n^{1/2}(\widehat{g}(\theta) - \mathrm{E}\widehat{g}(\theta)), \quad \text{where } \widehat{g}(\theta) := n^{-1}\sum_{i=1}^{n} g_i(\theta) \quad \text{and let}$$

$$\Delta(\theta) := \lim_{n\to\infty} \mathrm{E}\Psi_n(\theta)\Psi_n(\theta)' \in \mathbb{R}^{k\times k} \tag{2.3}$$

be the long-run covariance matrix of $g_i(\theta)$.[2] Let $\theta = (\alpha', \beta')'$, where $\alpha \in A$, $A \subset \mathbb{R}^{p_A}$, $\beta \in B$, $B \subset \mathbb{R}^{p_B}$, $\Theta = A \times B$, and $p_A + p_B = p$. The case $p_B = 0$ is allowed. In the following, we adopt Assumption C from

[2]Note that $\Delta(\theta)$, typically referred to as "long-run variance" in much of the econometrics literature, is proportional to the spectral density at zero frequency. Kernel-based spectral density estimation goes back to work by statisticians in the 1950s, see e.g. Parzen (1956, 1957) where consistency of spectral estimates is established for stationary time series, while studentization of mean-like statistics by a spectral density estimate goes back at least to Hannan (1957).

Stock and Wright (2000) in which α_0 and β_0 are modeled as weakly and strongly identified parameter vectors, respectively. For a detailed discussion of this assumption, see Stock and Wright (2000, pp. 1060–1061). Let $\mathcal{N} \subset B$ denote an open neighborhood β_0.[3]

Assumption ID. The true parameter $\theta_0 = (\alpha_0', \beta_0')'$ is in the interior of the compact set $\Theta = A \times B$ and (i) $\mathrm{E}\widehat{g}(\theta) = n^{-1/2}m_{1n}(\theta) + m_2(\beta)$, where $m_{1n}, m_1 : \Theta \to \mathbb{R}^k$ and (if $p_B > 0$) $m_2 : B \to \mathbb{R}^k$ are continuous functions such that $m_{1n}(\theta) \to m_1(\theta)$ uniformly on Θ, $m_1(\theta_0) = 0$ and $m_2(\beta) = 0$ if and only if $\beta = \beta_0$; (ii) $m_2 \in C^1(\mathcal{N})$; (iii) let $M_2(\beta) := (\partial m_2/\partial \beta)(\beta) \in \mathbb{R}^{k \times p_B}$. $M_2(\beta_0)$ has full column rank p_B.

Following the suggestion in Guggenberger (2003), we work with smoothed counterparts of the moment indicators $g_i(\theta)$ to handle the general time series setup considered here as in Kitamura and Stutzer (1997) and Smith (1997, 2001). See also Smith (2000, 2005) and Otsu (2006). An alternative procedure would be to work with a blocking method as in Kitamura (1997). For $i = 1, \ldots, n$, define

$$g_{in}(\theta) := S_n^{-1} \sum_{j=i-n}^{i-1} k(j/S_n) g_{i-j}(\theta), \tag{2.4}$$

where S_n is a bandwidth parameter ($S_n \to \infty$ as $n \to \infty$) and $k(\cdot)$ is a kernel. For simplicity, from now on the truncated kernel is used which is given by

$$k(x) = 1 \quad \text{if } |x| \leqslant 1 \quad \text{and} \quad k(x) = 0 \quad \text{otherwise} \tag{2.5}$$

and thus $g_{in}(\theta) = S_n^{-1} \sum_{j=\max\{-S_n, i-n\}}^{\min\{S_n, i-1\}} g_{i-j}(\theta)$.[4] Define

$$\widehat{g}_n(\theta) := n^{-1} \sum_{i=1}^{n} g_{in}(\theta) \quad \text{and} \quad \widehat{\Delta}(\theta) := S_n \sum_{i=1}^{n} g_{in}(\theta) g_{in}(\theta)'/n. \tag{2.6}$$

Under assumptions given in Lemma 2 below, the estimator $\widehat{\Delta}(\theta_0)$ is shown to be consistent for $2\Delta(\theta_0)$, whereas the "unsmoothed" version of the estimator, $\widehat{\Omega}(\theta_0)$, for

$$\widehat{\Omega}(\theta) := \sum_{i=1}^{n} g_i(\theta) g_i(\theta)'/n, \tag{2.7}$$

used in GS, while being consistent in an i.i.d. or m.d.s. setup, would not be consistent in the general time series context considered here. See GS's discussion of their assumption M_{θ_0}(ii). The consistency of $\widehat{\Delta}(\theta)$ is crucial for the testing procedures suggested in the next section. See Theorem 1 and Remark (2) below where we derive and discuss the asymptotic distribution of the test statistics in GS under the time series context considered here.

The statistics below are based on the GEL estimator. In what follows, a brief definition of the GEL estimator is given. For a more comprehensive discussion see Smith (1997, 2001), Newey and Smith (2004), and GS. Let ρ be a real-valued function $Q \to \mathbb{R}$, where Q is an open interval of the real line that contains 0 and

$$\widehat{\Lambda}_n(\theta) := \{\lambda \in \mathbb{R}^k : \lambda' g_{in}(\theta) \in Q \text{ for } i = 1, \ldots, n\}. \tag{2.8}$$

If defined, let $\rho_j(v) := (\partial^j \rho / \partial v^j)(v)$ and $\rho_j := \rho_j(0)$ for nonnegative integers j.

[3] Kleibergen (2005a, eq. (13), p. 1107) allows for a Jacobian matrix $J_\theta(\theta_0) := \lim_{n\to\infty} \mathrm{E}((\partial\widehat{g}(\theta)/\partial\theta)|_{\theta=\theta_0})$ (using our notation) that may or may not be of fixed full rank and may even equal zero (see the bottom of his p. 1108). Our Assumption ID can also account for this. For example, the case where $J_\theta(\theta_0) = 0$ corresponds to our setup with $\theta_0 = \alpha_0$ and neither β nor m_2 present. See (A.6) below.

[4] In general, one could employ kernels in the class $\mathcal{K}_1$ of Andrews (1991, p. 821) taking into account technical modifications in Jansson (2002); see, for example, Smith (2001) and Otsu (2006). Here we focus on the truncated kernel because it significantly simplifies the proofs and notation. In addition, for the testing purpose in this paper, it is not clear on what basis a kernel should be chosen and Monte Carlo simulations reveal that the finite-sample performance is not very sensitive to the kernel choice, see also Newey and West (1994) for similar findings in the HAC literature.

The GEL estimator is the solution to a saddle point problem

$$\widehat{\theta}_\rho := \arg\min_{\theta\in\Theta} \sup_{\lambda\in\widehat{\Lambda}_n(\theta)} \widehat{P}_\rho(\theta,\lambda), \quad \text{where} \tag{2.9}$$

$$\widehat{P}_\rho(\theta,\lambda) := 2\sum_{i=1}^{n}(\rho(\lambda' g_{in}(\theta)) - \rho_0)/n. \tag{2.10}$$

Assumption ρ. (i) ρ is concave on Q; (ii) ρ is C^2 in a neighborhood of 0 and $\rho_1 = \rho_2 = -1$.

Examples of GEL estimators include the CUE, see Hansen et al. (1996), empirical likelihood (EL, see Imbens, 1997; Qin and Lawless, 1994), and exponential tilting (ET, see Kitamura and Stutzer, 1997; Imbens et al., 1998) which correspond to $\rho(v) = -(1+v)^2/2$, $\rho(v) = \ln(1-v)$, and $\rho(v) = -\exp v$, respectively.

2.2. Test statistics

Here, statistics are introduced that can be used to test (2.2) in the time series model given by (2.1). It is established that they are asymptotically pivotal quantities and have limiting chi-square null distributions under Assumption ID. Therefore, these statistics lead to tests whose ERPs under the null should not be affected much by the strength or weakness of identification. There are other statistics that share this property in the general time series setup considered here, namely Kleibergen's (2005a) GMM-based and Otsu's (2006) GEL-based statistic. There are various other robust tests introduced for i.i.d. models, e.g. Kleibergen (2002), Caner (2003), and Moreira (2003). Kleibergen's and Otsu's statistics are compared to the approach of this paper in more detail below.

Let ρ be any function satisfying Assumption ρ. The first statistic is given by

$$GELR_\rho(\theta) := S_n^{-1} n\widehat{P}_\rho(\theta,\lambda(\theta))/2, \quad \text{where if it exists,}$$
$$\lambda(\theta) := \arg\max_{\lambda\in\widehat{\Lambda}_n(\theta)} \widehat{P}_\rho(\theta,\lambda). \tag{2.11}$$

The statistic $GELR_\rho(\theta)$ has a nonparametric likelihood ratio interpretation, see GS, where motivation is provided in the i.i.d. context. The generalization of the $GELR_\rho$ statistic in GS to the time series context has now been independently introduced by Otsu (2006), see his $\widehat{S}_{\mathrm{GEL}}$ statistic.

The second set of statistics is based on the FOC with respect to θ of the GEL estimator $\widehat{\theta}$. If the minimum of the objective function $\widehat{P}(\theta,\lambda(\theta))$ is obtained in the interior of Θ, the score vector with respect to θ must equal 0 at $\widehat{\theta}$. Using the envelope theorem it can be shown that this results in

$$0' = \lambda(\widehat{\theta})' \sum_{i=1}^{n} \rho_1(\lambda(\widehat{\theta})' g_{in}(\widehat{\theta}))G_{in}(\widehat{\theta})/n, \quad \text{where if defined} \tag{2.12}$$

$$G_{in}(\theta) := (\partial g_{in}/\partial\theta)(\theta) \in \mathbb{R}^{k\times p}; \tag{2.13}$$

see Newey and Smith (2004) and GS for a rigorous argument of this statement in the i.i.d. case. For $\theta\in\Theta$, define

$$D_\rho(\theta) := \sum_{i=1}^{n} \rho_1(\lambda(\theta)' g_{in}(\theta))G_{in}(\theta)/n \in \mathbb{R}^{k\times p}. \tag{2.14}$$

Thus, (2.12) may be written as $\lambda(\widehat{\theta})' D_\rho(\widehat{\theta}) = 0'$. The test statistic is given as a quadratic form in the score vector $\lambda(\theta)' D_\rho(\theta)$ evaluated at the hypothesized parameter vector θ and renormalized by the appropriate rate

$$S_\rho(\theta) := S_n^{-2} n\lambda(\theta)' D_\rho(\theta)(D_\rho(\theta)'\widehat{\Delta}(\theta)^{-1}D_\rho(\theta))^{-1}D_\rho(\theta)'\lambda(\theta)/2. \tag{2.15}$$

In addition, the following variant of $S_\rho(\theta)$:

$$LM_\rho(\theta) := n\widehat{g}_n(\theta)'\widehat{\Delta}(\theta)^{-1}D_\rho(\theta)(D_\rho(\theta)'\widehat{\Delta}(\theta)^{-1}D_\rho(\theta))^{-1}D_\rho(\theta)'\widehat{\Delta}(\theta)^{-1}\widehat{g}_n(\theta)/2 \tag{2.16}$$

is considered that substitutes $S_n^{-1}\lambda(\theta)$ in $S_\rho(\theta)$ by the asymptotically equivalent expression $-\widehat{\Delta}(\theta)^{-1}\widehat{g}_n(\theta)$, see Eq. (A.24) below. The names $S_\rho(\theta)$ and $LM_\rho(\theta)$ of the statistics are taken from GS and are based on the interpretation of the statistics as score and Lagrange multiplier statistics, respectively; see GS for more discussion. If $\rho(v) = \ln(1-v)$ we use the notation $LM_{\mathrm{EL}}(\theta)$ for $LM_\rho(\theta)$ and likewise for other statistics and functions ρ.

The next theorem discusses the asymptotic distribution of these test statistics evaluated at θ_0. To illustrate the need for smoothing, we also derive the asymptotic distribution of the test statistics in GS. In the following, the superscript "$*$" stands for unsmoothed expressions. Let $\widehat{\Delta}_n^*(\theta)$, $\widehat{P}_\rho^*(\theta,\lambda)$, $\lambda^*(\theta)$, and $D_\rho^*(\theta)$ be defined analogously to $\widehat{\Delta}_n(\theta)$, $\widehat{P}_\rho(\theta,\lambda)$, $\lambda(\theta)$, and $D_\rho(\theta)$ except that the smoothed expressions $g_{in}(\theta)$ and $G_{in}(\theta)$ are replaced by the unsmoothed expressions $g_i(\theta)$ and

$$G_i(\theta) := (\partial g_i/\partial\theta)(\theta). \tag{2.17}$$

The unsmoothed test statistics in GS corresponding to $GELR_\rho(\theta)$, $S_\rho(\theta)$, and $LM_\rho(\theta)$ can then be written as

$$\begin{aligned} &GELR_\rho^*(\theta) = n\widehat{P}_\rho^*(\theta,\lambda), \\ &S_\rho^*(\theta) = n\lambda^*(\theta)' D_\rho^*(\theta)(D_\rho^*(\theta)'\widehat{\Omega}(\theta)^{-1}D_\rho^*(\theta))^{-1}D_\rho^*(\theta)'\lambda^*(\theta) \quad \text{and} \\ &LM_\rho^*(\theta) = n\widehat{g}(\theta)'\widehat{\Omega}(\theta)^{-1}D_\rho^*(\theta)(D_\rho^*(\theta)'\widehat{\Omega}(\theta)^{-1}D_\rho^*(\theta))^{-1}D_\rho^*(\theta)'\widehat{\Omega}(\theta)^{-1}\widehat{g}(\theta). \end{aligned} \tag{2.18}$$

When deriving the asymptotic distribution of these statistics we assume that

$$\widehat{\Omega}(\theta_0)\to_{\mathrm{p}}\Omega(\theta_0) \quad \text{for } \Omega(\theta_0) := \lim_{n\to\infty} \mathrm{E}\sum_{i=1}^{n} g_i(\theta_0)g_i(\theta_0)'/n > 0. \tag{2.19}$$

The technical assumptions M_{θ_0} and their interpretation are given in the Appendix.

Theorem 1. *Suppose ID, ρ, and M_{θ_0}(i)–(iii) hold. Then for $S_n\to\infty$ as $n\to\infty$ and $S_n = o(n^{1/2})$ it follows that:*

(i) $GELR_\rho(\theta_0)\to_{\mathrm{d}}\chi^2(k)$ *and* (i)$'$ $GELR_\rho^*(\theta_0)\to_{\mathrm{d}}\xi'\Omega(\theta_0)^{-1}\xi$,

where ξ is a random vector distributed as $\mathrm{N}(0,\Delta(\theta_0))$. *If in addition* M_{θ_0} (iv)–(vii) *hold then*

(ii) $S_\rho(\theta_0), LM_\rho(\theta_0)\to_{\mathrm{d}}\chi^2(p)$ *and* (ii)$'$ $S_\rho^*(\theta_0), LM_\rho^*(\theta_0)\to_{\mathrm{d}}\widetilde{\xi}$,

where $\widetilde{\xi}$ is a random variable defined in (A.36) *in the Appendix and where for the unsmoothed statistics we assume* (2.19) *and the analogous formula* (A.10) *for derivatives of $g_i(\theta_0)$.*

Remarks. (1) Theorem 1 implies a straightforward method to construct confidence regions or hypothesis tests for θ_0 based on the smoothed statistics. For example, a critical region for test (2.2) at significance level r is given by $\{GELR_\rho(\theta_0)\geq\chi_r^2(k)\}$, where $\chi_r^2(k)$ denotes the $(1-r)$-critical value from the $\chi^2(k)$ distribution. In contrast to classical test statistics such as a Wald statistic, the statistics $GELR_\rho(\theta_0)$, $S_\rho(\theta_0)$, and $LM_\rho(\theta_0)$ are asymptotically pivotal statistics under Assumption ID. Therefore, ERPs under the null of tests based on these statistics should not vary much with the strength or weakness of identification in finite samples. For the statistics $S_\rho(\theta_0)$ and $LM_\rho(\theta_0)$ to be pivotal, it is crucial that $D_\rho(\theta_0)$ (appropriately renormalized) and $n^{1/2}\widehat{g}_n(\theta_0)$ are asymptotically independent under both weak and strong identification, see the proof of the theorem. Also see Smith (2001) which demonstrates this property for the strongly identified case. Theorem 1 also shows that the asymptotic null distribution of the test statistics does not depend on the choice of ρ.

(2) Theorem 1(i)$'$ and (ii)$'$ shows that in the general time series context considered here, smoothing of the moment conditions is necessary to obtain test statistics whose asymptotic distributions are nuisance parameter free. While $n^{1/2}\widehat{g}_n(\theta_0)$ and $n^{1/2}\widehat{g}(\theta_0)$ differ only by a proportionality factor (see Lemma 1), the crucial consequence of smoothing is that the (renormalized) quantities $\widehat{g}_n(\theta_0)$ and $D_\rho(\theta_0)$ are asymptotically independent while their unsmoothed counterparts $\widehat{g}(\theta_0)$ and $D_\rho^*(\theta_0)$ are not. See Eqs. (A.29), (A.34), and subsequent analysis in the Appendix. Another important result of smoothing is that the estimator $\widehat{\Delta}(\theta_0)$ is consistent for $2\Delta(\theta_0)$ while the unsmoothed counterpart $\widehat{\Omega}(\theta_0)$ is generally inconsistent.

Important recent work by Kiefer et al. (2000) and Kiefer and Vogelsang (2002a,b,2005) shows that in regression models with correlated errors a t- or F-test can be successfully implemented without using a consistent HAC estimator of the long-run variance matrix of the parameter estimator. They use (inconsistent) variance estimators—implemented with bandwidth b equal to the sample size or equal to a fixed portion of the sample size—that converge to a limiting random matrix that is proportional to the long-run variance matrix. They show that their test statistics converge in distribution to nuisance parameter free functionals of a Wiener process. In their model there are no (weak) instruments as in ours, but even with strong instruments (i.e. $p_A = 0$) (A.36) shows that generally $LM^*_\rho(\theta_0) \not\to_d \chi^2(p)$ because generally $\Omega(\theta_0)$ and $\Delta(\theta_0)$ differ.

We now consider a simple example to show that in our model we do not obtain a nuisance parameter free distribution if we use the unsmoothed statistics. We focus on $GELR^*_\rho(\theta_0)$ whose limit distribution is given by $\xi'\Omega(\theta_0)^{-1}\xi$. Consider the linear IV regression model given in (3.1) below where for simplicity we assume that there is only one instrument and, as described in Section 3.1, u_i and Z_i are independent zero mean AR(1) processes with autoregressive (AR) parameter equal to ϕ_u and ϕ_Z, respectively. Then,

$$\begin{aligned}
\Omega(\theta_0) &= \lim_{n\to\infty} \sum_{i=1}^{n} \mathrm{E}u_i^2 Z_i^2/n = \mathrm{E}u_i^2 \mathrm{E}Z_i^2 = (1-\phi_u^2)^{-1}(1-\phi_Z^2)^{-1},\\
\Delta(\theta_0) &= \lim_{n\to\infty} n\mathrm{E}\widehat{g}(\theta_0)^2 = \lim_{n\to\infty} \sum_{i,j=1}^{n} \mathrm{E}u_i u_j \mathrm{E}Z_i Z_j/n\\
&= \Omega(\theta_0)\left[1 + 2\lim_{n\to\infty} \sum_{i=1}^{n-1} \frac{n-i}{n}(\phi_u\phi_Z)^i\right]\\
&= \Omega(\theta_0)[1 + 2\phi_u\phi_Z(1-\phi_u\phi_Z)^{-1}];
\end{aligned} \tag{2.20}$$

see e.g. Hamilton (1994, p. 53).[5] Therefore, $GELR^*_\rho(\theta_0) \to_d c\chi^2(1)$, where $c := 1 + 2\phi_u\phi_Z(1-\phi_u\phi_Z)^{-1}$. A test based on $GELR^*_\rho$ using $\chi^2(1)$ critical values overrejects (underrejects) under the null if $c>1$ ($c<1$). Opposing values of ϕ_u and ϕ_Z reduce the degree of overrejection and can even lead to underrejection. Our Monte Carlo study below finds that this property seems to hold more generally for all the statistics considered in this paper. For example, for $\phi_u = \phi_Z$ we have $c = 1 + 2\phi_u^2(1-\phi_u^2)^{-1}$ while for $\phi_u = -\phi_Z$ we have $c = 1 - 2\phi_u^2(1+\phi_u^2)^{-1}$. While the latter quantity is always smaller than 1 for $\phi_u \neq 0$ and converges to 0 for $|\phi_u| \to 1$, the former quantity is always bigger than 1 for $\phi_u \neq 0$ and diverges to $+\infty$ for $|\phi_u| \to 1$.

(3) A drawback of $GELR_\rho(\theta_0)$ is that its limiting null distribution has degrees of freedom equal to k, the number of moment conditions rather than the dimension of the parameter vector p. In general, this has a negative impact on the power properties of hypothesis tests based on $GELR_\rho(\theta_0)$ in over-identified situations. On the other hand, the limiting null distribution of $S_\rho(\theta_0)$ and $LM_\rho(\theta_0)$ has degrees of freedom equal to p. Therefore, the power of tests based on these statistics should not be negatively affected by a high degree of over-identification.

(4) Assumption M_{θ_0} (given in the Appendix) is compatible with many time series models and, besides technicalities, essentially states (i) that the Bartlett HAC estimator consistently estimates the long-run variance matrix $\Delta(\theta_0)$ and (ii) that a CLT holds for the times series $(vec\, G'_{iA}(\theta_0), g'_i(\theta_0))'$ with full rank asymptotic covariance matrix $V(\theta_0)$, where $G_{iA}(\theta_0)$ is the submatrix of $G_i(\theta_0)$ corresponding to the weakly identified parameters, see the Appendix for a detailed discussion. Part (ii) is very closely related to Assumption 1 in Kleibergen (2005a) that states a CLT for $(vec\, G'_i(\theta_0), g'_i(\theta_0))'$ with possibly singular covariance matrix. Therefore, the approach taken in this paper generalizes the setup in GS whose applications were restricted to m.d.s.

(5) The theorem does not give any guidelines on how to choose the bandwidth S_n in finite samples. In fact, just as for the choice of the kernel k, it is difficult to provide theory for its choice in the testing context considered here, where size and power properties matter. One could still follow Andrews (1991) and choose S_n such that the mean-squared error of the covariance matrix estimator is minimized after a time series model has

[5] Note that for any $q \in \mathbb{R}$ we have $\sum_{i=1}^{n-1}((n-i)/n)q^i = ((1-q)(q+n^{-1}) - n^{-1}(1-q^{n+1}))/(1-q)^2$ which for $|q|<1$ converges to $q/(1-q)$ as $n \to \infty$.

been specified. However, it is unclear what effect this procedure would have on size and power of the test and it would be surprising if this procedure led to any optimality property.

2.3. Sub-vector statistics

We are now interested in testing

$$\mathrm{H}_0\colon \alpha_0 = \overline{\alpha} \quad \text{versus} \quad \mathrm{H}_1\colon \alpha_0 \neq \overline{\alpha}, \tag{2.21}$$

where $\alpha_0 \in \mathbb{R}^{p_A}$ and $\theta_0 = (\alpha_0', \beta_0')'$. Let $\theta = (\alpha_1', \alpha_2', \beta')'$, where $\alpha_j \in A_j$, $A_j \subset \mathbb{R}^{p_{A_j}}$ $(j = 1, 2)$, $p_{A_1} + p_{A_2} = p_A$, and $\beta \in B$, $B \subset \mathbb{R}^{p_B}$. We assume that the true parameter $\theta_0 = (\alpha_{01}', \alpha_{02}', \beta_0')'$ is in the interior of the compact space Θ, where $\Theta = A_1 \times A_2 \times B$. We now modify Assumption ID. Let $\mathcal{N} \subset A_2 \times B$ be an open neighborhood of (α_{02}, β_0).

Assumption $\mathbf{ID}_{\alpha_0}$. (i) $E\widehat{g}(\theta) = n^{-1/2} m_{1n}(\theta) + m_2(\alpha_2, \beta)$, where $m_{1n}, m_1\colon \Theta \to \mathbb{R}^k$ and (if $p_{A_2} + p_B > 0$) $m_2\colon A_2 \times B \to \mathbb{R}^k$ are continuous functions such that $m_{1n}(\theta) \to m_1(\theta)$ uniformly on Θ, $m_1(\theta_0) = 0$ and $m_2(\alpha_2, \beta) = 0$ if and only if $(\alpha_2, \beta) = (\alpha_{02}, \beta_0)$; (ii) $m_2 \in C^1(\mathcal{N})$; (iii) let $M_2(\cdot) := (\partial m_2/\partial(\alpha_2', \beta')')(\cdot) \in \mathbb{R}^{k \times (p_{A_2} + p_B)}$. $M_2(\alpha_{02}, \beta_0)$ has full column rank $p_{A_2} + p_B$.[6]

Assumption ID_{α_0} implies that α_{01} is weakly and (α_{02}, β_0) is strongly identified. To adapt the full-vector test statistics to the sub-vector case, the basic idea is to replace β by an estimator $\widehat{\beta}(\alpha)$. Define the GEL estimator $\widehat{\beta}(\alpha)$ for β_0:

$$\widehat{\beta}(\alpha) := \arg\min_{\beta \in B} \sup_{\lambda \in \widehat{\Lambda}_n(\alpha', \beta')'} \widehat{P}((\alpha', \beta')', \lambda). \tag{2.22}$$

Our assumptions below imply consistency $\widehat{\beta} := \widehat{\beta}(\alpha_0) \to_p \beta_0$ and efficiency under the null hypothesis: also see Smith (2001). Let

$$\widehat{\theta}_0 := (\alpha_0', \widehat{\beta}(\alpha_0)')' \quad \text{and} \quad \theta_\beta := (\alpha_0', \beta')'. \tag{2.23}$$

We now introduce the sub-vector statistics. Recall the definition of $GELR_\rho(\theta)$ in (2.11). Evaluated at $\alpha = \alpha_0$, the $GELR_\rho$ sub-vector test statistic is given by

$$GELR_\rho^{\mathrm{sub}}(\alpha_0) := GELR_\rho(\widehat{\theta}_0). \tag{2.24}$$

We now generalize the statistics S_ρ and LM_ρ to the sub-vector case. The motivation of these statistics is analogous to the sub-vector statistics in GS. We need additional notation. For a full column rank matrix $A \in \mathbb{R}^{k \times p}$ and $0 < K \in \mathbb{R}^{k \times k}$, let $P_A(K) := A(A'K^{-1}A)^{-1}A'K^{-1}$ and $M_A(K) := I_k - P_A(K)$. We abbreviate this notation to P_A and M_A if $K = I_k$. If $p = 0$, set $M_A = I_k$. Let

$$D_\rho(\alpha_0) := \sum_{i=1}^n \rho_1(\lambda(\widehat{\theta}_0)' g_{in}(\widehat{\theta}_0)) G_{inA}(\widehat{\theta}_0)/n \in \mathbb{R}^{k \times p_A}, \tag{2.25}$$

where $G_{inA}(\theta)$ is defined by $G_{in}(\theta) = (G_{inA}(\theta), G_{inB}(\theta))$ for $G_{inA}(\theta) \in \mathbb{R}^{k \times p_A}$ and $G_{inB}(\theta) \in \mathbb{R}^{k \times p_B}$; see Eq. (2.13). The definition of $D_\rho(\alpha_0)$ coincides with the one of $D_\rho(\theta_0)$ when α_0 is the full vector θ_0. If $p_B > 0$ let

$$\widehat{M}(\alpha_0) := \widehat{\Delta}(\widehat{\theta}_0)^{-1} M_{\widehat{G}_B(\widehat{\theta}_0)}(\widehat{\Delta}(\widehat{\theta}_0)/2), \tag{2.26}$$

and otherwise let $\widehat{M}(\alpha_0) := \widehat{\Delta}(\widehat{\theta}_0)^{-1}$, where

$$\widehat{G}(\theta) := n^{-1} \sum_{i=1}^n G_i(\theta) \in \mathbb{R}^{k \times p}, \quad \widehat{G}(\theta) = (\widehat{G}_A(\theta), \widehat{G}_B(\theta)) \tag{2.27}$$

[6] In this subsection, $m_2(\cdot)$ and $M_2(\cdot)$ (defined already in ID_{θ_0} above as functions of β) now denote functions of α_2 and β.

for $\widehat{G}_A(\theta) \in \mathbb{R}^{k\times p_A}$ and $\widehat{G}_B(\theta) \in \mathbb{R}^{k\times p_B}$. The sub-vector test statistic $S_\rho^{\text{sub}}(\alpha_0)$ is constructed as a quadratic form in the vector of FOC $\lambda(\widehat{\theta}_0)'D_\rho(\widehat{\theta}_0)$ with weighting matrix given by $\widehat{M}(\alpha_0)$. Let

$$S_\rho^{\text{sub}}(\alpha_0) := nS_n^{-2}\lambda(\widehat{\theta}_0)'D_\rho(\alpha_0)(D_\rho(\alpha_0)'\widehat{M}(\alpha_0)D_\rho(\alpha_0))^{-1}D_\rho(\alpha_0)'\lambda(\widehat{\theta}_0)/2. \tag{2.28}$$

The statistic $LM_\rho^{\text{sub}}(\alpha_0)$ is constructed like $S_\rho^{\text{sub}}(\alpha_0)$ but replaces $n^{1/2}S_n^{-1}\lambda(\widehat{\theta}_0)$ by the asymptotically equivalent expression $-\widehat{\Delta}(\widehat{\theta}_0)^{-1}n^{1/2}\widehat{g}_n(\widehat{\theta}_0)$. Therefore,

$$LM_\rho^{\text{sub}}(\alpha_0) := n\widehat{g}_n(\widehat{\theta}_0)'\widehat{\Delta}(\widehat{\theta}_0)^{-1}D_\rho(\alpha_0)(D_\rho(\alpha_0)'\widehat{M}(\alpha_0)D_\rho(\alpha_0))^{-1}D_\rho(\alpha_0)'\widehat{\Delta}(\widehat{\theta}_0)^{-1}\widehat{g}_n(\widehat{\theta}_0)/2. \tag{2.29}$$

Under Assumption M_{α_0} given in the Appendix we have the following theorem.[7]

Theorem 2. (i) *Assume* $1 \leqslant p_A < p$. *Suppose Assumptions* ID_{α_0}, M_{α_0}(i)–(iv), *and* ρ *hold. Then*,

$$GELR_\rho^{\text{sub}}(\alpha_0) \to_{\text{d}} \chi^2(k - p_B).$$

(ii) *If in addition* M_{α_0}(v)–(vii) *hold*, *then*

$$S_\rho^{\text{sub}}(\alpha_0) \quad \text{and} \quad LM_\rho^{\text{sub}}(\alpha_0) \to_{\text{d}} \chi^2(p_A).$$

Under the assumption used here, that the parameters not under test are strongly identified, there are various other alternatives for sub-vector inference besides $GELR_\rho^{\text{sub}}(\alpha_0)$, $S_\rho^{\text{sub}}(\alpha_0)$, and $LM_\rho^{\text{sub}}(\alpha_0)$. See, for example, the tests by Kleibergen (2004, 2005a) and Otsu (2006). An interesting recent contribution by Kleibergen (2005b) introduces boundedly pivotal tests for the linear IV model *without* additional identification assumptions. Alternatively, confidence intervals can be constructed by a projection argument; see Dufour (1997). However, this approach is conservative and in general computationally cumbersome. In a recent paper, Dufour and Taamouti (2005) show that the Anderson and Rubin (1949) statistic is an exception, in that a closed form solution is available. Another alternative is Guggenberger and Wolf (2004) who suggest a subsampling approach. In contrast to some of the above procedures, subsampling leads to sub-vector tests whose null rejection probability converges to the desired nominal level *without* additional identification assumptions for each *fixed* degree of identification. Guggenberger and Wolf's (2004) Monte Carlos suggest that for sub-vector inference subsampling seems to do better in terms of power than Kleibergen (2004, 2005a) and Dufour and Taamouti (2005). In their simulation study, the former procedure tends to underreject when the components not under test are only weakly identified and the latter seems to underreject across all the scenarios. On the other hand, they find that for full-vector inference, subsampling is outperformed by the procedures in GS and Kleibergen (2005a). Andrews and Guggenberger's (2005b,c) size correction methods for subsampling tests could also be applied to sub-vector tests.

2.4. Comparison with Kleibergen (2005a) and Otsu (2006)

Here, we compare our (full-vector) statistics to the K and $\widehat{K}_{\text{GEL}}$ statistics of Kleibergen (2005a) and Otsu (2006). These statistics, S_ρ and LM_ρ, and the ones defined below have the same first-order theory under the null hypothesis; asymptotically they are all distributed as $\chi^2(p)$ under the null.

Kleibergen's K statistic is defined as

$$K(\theta) := n\widehat{g}(\theta)'\widetilde{\Delta}(\theta)^{-1}D_\theta(D_\theta'\widetilde{\Delta}(\theta)^{-1}D_\theta)^{-1}D_\theta'\widetilde{\Delta}(\theta)^{-1}\widehat{g}(\theta), \quad \text{where}$$
$$D_\theta := \widehat{G}(\theta) - \widetilde{\Omega}(\theta)[I_p \otimes (\widetilde{\Delta}(\theta)^{-1}\widehat{g}(\theta))] \in \mathbb{R}^{k\times p} \quad \text{and} \tag{2.30}$$

$\widetilde{\Delta}(\theta)$ and $\widetilde{\Omega}(\theta)$ are consistent estimators for $\Delta(\theta)$ and the long-run covariance matrix $\lim_{n\to\infty} \mathrm{E}\{n^{-1}\sum_{i,j=1}^n [G_i(\theta) - \mathrm{E}G_i(\theta)][(I_p \otimes g_j(\theta)') - \mathrm{E}(I_p \otimes g_j(\theta)')]\}$, respectively. Kleibergen (2005a) suggests the use of HAC estimators for $\widetilde{\Delta}(\theta)$ and $\widetilde{\Omega}(\theta)$; see e.g. Andrews (1991). The statistics LM_ρ and the K statistic are given as quadratic forms in the FOC of the GEL and the GMM CUE estimator, respectively. The intuition for tests

[7]Note that there is a typo in GS (p. 685, last line): p_A should be p_B.

based on these statistics is as follows: under strong identification, GEL and GMM estimators are consistent. In consequence, in large samples the FOC for the estimator also holds at the true parameter vector θ_0. Therefore, the statistics are quadratic forms which are expected to be small at the true vector θ_0. Even though the GMM CUE and GEL CUE are numerically identical (see Newey and Smith, 2004, footnote 2), their FOC are different and therefore LM_{CUE} and K will typically differ. For i.i.d. or m.d.s. scenarios GS specify for which estimators $\widetilde{\Delta}(\theta)$ and $\widetilde{\Omega}(\theta)$ in the K statistic, K and LM_{CUE} are identical. These statements in GS cannot be generalized to the general time series setup, where K and LM_{CUE} are different. One reason for that is that in this latter statistic functions of the smoothed indicators g_{in} and G_{in} are used, e.g. $\widehat{g}_n$, while the former statistic uses functions of the unsmoothed indicators, e.g. $\widehat{g}$.

To assess which factor in LM_ρ accounts for most of the finite-sample differences between $K(\theta)$ and LM_ρ we also consider the following hybrid statistics $K_{\rho,H_j}(\theta)$ in our Monte Carlo study below. $K_{\rho,H_1}(\theta)$ replaces $\widehat{g}_n(\theta)$ in LM_ρ by $2\widehat{g}(\theta)$[8];

$$K_{\rho,H_1}(\theta):=2n\widehat{g}(\theta)'\widehat{\Delta}(\theta)^{-1}D_\rho(\theta)(D_\rho(\theta)'\widehat{\Delta}(\theta)^{-1}D_\rho(\theta))^{-1}D_\rho(\theta)'\widehat{\Delta}(\theta)^{-1}\widehat{g}(\theta). \tag{2.31}$$

$K_{\rho,H_2}(\theta)$ replaces $\widehat{\Delta}(\theta)$ in $K_{\rho,H_1}(\theta)$ by $2\widetilde{\Delta}(\theta)$ (where $\widetilde{\Delta}(\theta_0)\to_{\mathrm{p}}\Delta(\theta_0)$ is a HAC estimator):

$$K_{\rho,H_2}(\theta):=n\widehat{g}(\theta)'\widetilde{\Delta}(\theta)^{-1}D_\rho(\theta)(D_\rho(\theta)'\widetilde{\Delta}(\theta)^{-1}D_\rho(\theta))^{-1}D_\rho(\theta)'\widetilde{\Delta}(\theta)^{-1}\widehat{g}(\theta). \tag{2.32}$$

By Lemma 1 below these changes do not affect the limit distribution, and, as for LM_ρ, we have $K_{\rho,H_j}(\theta_0)\to_{\mathrm{d}}\chi^2(p)$ for $j=1,2$. Kleibergen's (2005a) statistic $K(\theta)$ and the hybrid statistic $K_{\rho,H_2}(\theta)$ only differ by the choice of the matrix D_θ and $D_\rho(\theta)$, respectively.

Otsu's (2006) statistic is given by

$$\widehat{K}_{\mathrm{GEL}}(\theta):=S_n^{-1}n\sup_{\gamma\in\Gamma(\theta)}\widehat{P}_\rho(\theta,\widehat{\Delta}(\theta)^{-1}D_\rho(\theta)\gamma)/2, \quad \text{where}$$
$$\Gamma(\theta):=\{\gamma\in\mathbb{R}^p;\widehat{\Delta}(\theta)^{-1}D_\rho(\theta)\gamma\in\widehat{\Lambda}_n(\theta)\} \quad \text{and} \tag{2.33}$$

$\widehat{\Delta}(\theta)$ and $D_\rho(\theta)$ are defined in (2.6) and (2.14), respectively. Here, $\widehat{K}_{\mathrm{GEL}}(\theta)$ has been formulated based on the truncated kernel but can of course be implemented using more general kernels, see Otsu (2006); also instead of $\widehat{\Delta}(\theta)$, any other consistent covariance matrix estimator could be used. $\widehat{K}_{\mathrm{GEL}}(\theta)$ is not given as a quadratic form in the FOC and the above intuition does not apply. In contrast to the $GELR_\rho$ statistic, however, the asymptotic null distribution of $\widehat{K}_{\mathrm{GEL}}$ does not depend on the number of moment conditions k. This is achieved by considering the transformed moment indicators $g_{in}'\widehat{\Delta}(\theta)^{-1}D_\rho(\theta)$ in (2.33) rather than g_{in}' as in (2.11). A drawback of Otsu's (2006) approach is that two maximizations are necessary to calculate the statistic, one to calculate $\lambda(\theta)$ in $D_\rho(\theta)$ of (2.14) and one in (2.33). The latter maximization may be simply avoided as follows. Define k-vectors

$$\mu_\rho(\theta):=-S_n\widehat{\Delta}(\theta)^{-1}D_\rho(\theta)(D_\rho(\theta)'\widehat{\Delta}(\theta)^{-1}D_\rho(\theta))^{-1}D_\rho(\theta)'\widehat{\Delta}(\theta)^{-1}\widehat{g}_n(\theta),$$
$$\tilde{\mu}_\rho(\theta):=\widehat{\Delta}(\theta)^{-1}D_\rho(\theta)(D_\rho(\theta)'\widehat{\Delta}(\theta)^{-1}D_\rho(\theta))^{-1}D_\rho(\theta)'\lambda(\theta). \tag{2.34}$$

Define the statistic

$$GELR_\rho(\theta,\mu):=S_n^{-1}n\widehat{P}_\rho(\theta,\mu)/2. \tag{2.35}$$

Theorem 3. *Suppose ID, ρ, and M_{θ_0}(i)–(vii) hold. Then for $S_n\to\infty$ as $n\to\infty$ and $S_n=\mathrm{o}(n^{1/2})$ it follows that*

$$GELR_\rho(\theta_0,\mu_\rho(\theta_0)),\, GELR_\rho(\theta_0,\tilde{\mu}_\rho(\theta_0))\to_{\mathrm{d}}\chi^2(p).$$

Remark. The function ρ used in obtaining $\mu_\rho(\theta)$ or $\tilde{\mu}_\rho(\theta)$ through $D_\rho(\theta)$ and $\lambda(\theta)$ may be allowed to differ from that defining $GELR_\rho(\theta,\mu)$ as long as both functions satisfy Assumption ρ. Note that even though the statistics in Theorem 3 are first-order equivalent to Otsu's (2006) $\widehat{K}_{\mathrm{GEL}}(\theta_0)$ test statistic, they are in general not numerically equal. We compare their performance in the Monte Carlo study in the next section.

[8]We would like to thank a referee for suggesting these hybrid statistics. The Monte Carlo study below indicates that $K_{\rho,H_1}(\theta)$ has very favorable size and $K_{\rho,H_2}(\theta)$ has very favorable power properties.

3. Monte Carlo study

In this section, the finite-sample properties of the hypotheses tests in Theorems 1 and 3 are investigated in a Monte Carlo study and compared to the tests suggested in Kleibergen (2005a) and Otsu (2006). To better understand the performance differences between LM_ρ and K, we also include the hybrid statistics K_{ρ,H_j} for $j = 1, 2$ defined in (2.31) and (2.32) in our study.

3.1. Design

The data generating process is given by the linear IV time series model

$$\begin{aligned} y &= Y\theta_0 + u, \\ Y &= Z\Pi + V. \end{aligned} \tag{3.1}$$

There is only a single right-hand side endogenous variable Y and no included exogenous variables. Let $Z \in \mathbb{R}^{n\times k}$, where k is the number of instruments and n the sample size. The reduced form matrix $\Pi \in \mathbb{R}^k$ equals a vector of ones times a constant Π_1 that determines the strength or weakness of identification. Similar to the design in Otsu (2006), each column of Z and u is generated as zero mean AR(1) or MA(1) processes (with AR and moving-average (MA) parameters ϕ and v, respectively) with innovations distributed as independent N(0,1) random variables and V has i.i.d. N(0,1) components. To generate an AR(1) process, $u_i = \phi u_{i-1} + \varepsilon_i$ say, we set $u_0 = 0$. MA(1) processes $\{u_i\}$ with MA parameter v are generated as $u_i = \varepsilon_i - v\varepsilon_{i-1}$. The innovations of the process for u, ε_i say, and the ith component of V are correlated; their joint distribution is $N(0, \Sigma)$, where $\Sigma \in \mathbb{R}^{2\times 2}$ with diagonal elements equal to unity and off-diagonal elements ρ_{uV}.

Interest focuses on testing the scalar null hypothesis $H_0\colon \theta_0 = 0$ versus the alternative hypothesis $H_1\colon \theta_0 \neq 0$. Results are reported at nominal levels of 5% for sample size $n = 200$. The following 60 parameter combinations are considered. Twelve combinations of k, Π_1, and ρ_{uV}:

$$k = 2, 10, 20, \quad \Pi_1 = .01, .5, \quad \rho_{uV} = 0, .5 \tag{3.2}$$

times the five AR(1)/MA(1) specifications

$$\phi = 0, .5, .9, \quad v = .5, .9 \tag{3.3}$$

are considered. We also consider an additional $12 \times 4 = 48$ parameter combinations where this time the AR/MA parameter for the AR(1) or MA(1) processes in the columns of Z equals -1 times the AR/MA parameter for the AR(1) or MA(1) process u and the latter parameter takes on the values $\phi = .5, .9$ or $v = .5, .9$. We call these cases designs with "opposing" AR/MA parameters whereas the other cases are called designs with "same" AR/MA parameters.

We report results for the seven statistics LM_{EL}, K_{EL,H_j}, for $j = 1, 2$, $GELR_{\mathrm{ET}}(\theta_0, \mu_{\mathrm{EL}}(\theta_0))$, $GELR_{\mathrm{ET}}(\theta_0, \tilde{\mu}_{\mathrm{EL}}(\theta_0))$, $\widehat{K}_{\mathrm{GEL}}$, and K in the study.[9] For K_{ρ,H_2} and K we use a Bartlett kernel to calculate the covariance matrix estimators and for $\widehat{K}_{\mathrm{GEL}}$ we use the EL specification. We use the ET specification for the statistics from Theorem 3 because in finite samples $1 - \mu_{\mathrm{EL}}(\theta_0)' g_{in}(\theta_0)$ or $1 - \tilde{\mu}_{\mathrm{EL}}(\theta_0)' g_{in}(\theta_0)$ is sometimes negative which prevents us from calculating the EL criterion function.

To implement the statistics, the bandwidth S_n has to be chosen. We consider fixed bandwidths $S_n = 1, \ldots, 15$ and also calculate the i.i.d. versions of the test statistics. Note that for the Bartlett kernel, $S_n = 1$ leads to numerically identical results for K as no smoothing. To solve the maximization problems in λ in the GEL-based statistics, a Newton–Raphson algorithm is used. Size and power properties are investigated by considering $\theta_0 = 0, 1$, and -1. All results reported below are based on 20,000 simulation repetitions.

[9] We also included five additional tests based on S_{EL}, $GELR_{\mathrm{EL}}$, a variant of K that uses a recentered HAC estimator (as suggested in Kleibergen, 2005a, p. 1112), and two variants of $\widehat{K}_{\mathrm{GEL}}$ that use a recentered Bartlett HAC estimator and a Bartlett HAC estimator. The first four of those tests have far less desirable size properties in our study across virtually all designs and bandwidths than all the other tests and the last test is dominated in terms of size by $\widehat{K}_{\mathrm{GEL}}$. Therefore, detailed results for those tests are not discussed here. Size problems of the i.i.d. versions of S_{EL} and $GELR_{\mathrm{EL}}$ in finite samples were also reported in GS.

3.2. Results

There are various patterns in our simulation results that allow us to restrict our discussion to a certain subset of the many designs:

As to be expected, all the tests have reliable size properties for the i.i.d. case $\phi = 0$ without any smoothing. For each of the test statistics the ERPs under the null are very similar for the three cases of same AR/MA parameters when $\phi = .5$, $v = .5, .9$, whereas the case $\phi = .9$ is characterized by uniformly much higher ERPs. Furthermore, opposing values of the AR or MA parameters in the u and Z processes typically lead to ERPs under the null that are—with few exceptions—uniformly smaller (or equal) than the nominal size across all test statistics, bandwidth choices S_n, and parameter combinations, and ERPs are smallest when $\phi = .9$. This generalizes our findings in Remark (2) above from the unsmoothed statistic $GELR^*_\rho$ to all the statistics considered in this study. In sum, our discussion of potential size distortion of the testing procedures can be reduced to the AR(1) cases $\phi = .5$ and .9 where the AR parameters of u and Z have the same sign. The power results for $\theta_0 = -1$ and 1 are qualitatively very similar and therefore we restrict attention to the former. Furthermore, power results for all AR/MA cases are virtually identical for almost all cases and statistics except for the case of same AR parameter when $\phi = .9$; therefore, as for size, we can w.l.o.g. restrict our discussion to the two cases of same AR parameter $\phi = .5$ and .9.

The ERPs under the null and alternative are qualitatively identical for the two cases $\rho_{uV} = 0, .5$ across all statistics and almost all designs[10] and thus, in what follows, we restrict attention to $\rho_{uV} = .5$.

As to be expected from our theory, the ERPs under the null do generally not vary much with Π_1, the strength of the instruments. Therefore regarding size properties, we restrict the following discussion to the case $\Pi_1 = .01$. In contrast to size, power properties do of course strongly depend on Π_1, with ERPs of the test statistics being often in the close vicinity of the nominal size of the test when $\Pi_1 = .01$.[11] Therefore, for power, we restrict attention to $\Pi_1 = .5$.

Based on the above discussion we select the following figures. Figs. 1(1– 4) contain size and Figs. 2 (1–3), 3(1–3) power curves of the LM_{EL}, K_{EL,H_j}, for $j = 1, 2$, $GELR_{\mathrm{ET}}(\theta_0, \mu_{\mathrm{EL}}(\theta_0))$, $GELR_{\mathrm{ET}}(\theta_0, \tilde{\mu}_{\mathrm{EL}}(\theta_0))$ (referred to as $GELR_1$ and $GELR_2$ in the figures), $\widehat{K}_{\mathrm{GEL}}$, and K tests as functions of the bandwidths $S_n = 1, \ldots, 15$ for the cases

$$\begin{aligned}
&\text{Fig. 1:} \quad k = 2, 20,\ \Pi_1 = .01,\ \rho_{uV} = .5,\ \phi = .5, .9,\ \theta_0 = 0 \quad \text{(size)},\\
&\text{Fig. 2:} \quad k = 2, 10, 20,\ \Pi_1 = .5,\ \rho_{uV} = .5,\ \phi = .5,\ \theta_0 = -1 \quad \text{(power)},\\
&\text{Fig. 3:} \quad k = 2, 10, 20,\ \Pi_1 = .5,\ \rho_{uV} = .5,\ \phi = .9,\ \theta_0 = -1 \quad \text{(power)},
\end{aligned} \tag{3.4}$$

where the processes u and Z have AR parameters of the same sign. For convenience, at $S_n = 0$ we report the results for the unsmoothed i.i.d. versions of the statistics. Since we do not provide a data driven method of choosing S_n, we report results for an array of S_n values. To interpret the figures, as long as there is no data driven choice for S_n, it is desirable for a testing procedure to have ERPs under the null that come close to the nominal size for a wide array of bandwidth choices; in other words, little dependence of the performance of the test on the choice of the bandwidth is desired.

All results not reported here are available from the authors upon request.

We now discuss the size and power results in more detail using the above figures as guiding examples.

We first discuss the *size* results. As to be expected from Theorem 1, all tests are typically size-distorted in the time series models with same AR/MA parameters when there is no smoothing. Typically, the higher the AR coefficient ϕ the higher the size distortion, e.g. compare Figs. 1(1 and 2) to 1(3 and 4), respectively. On the other hand, as to be expected from Remark (2) above, for opposing values of the AR parameter the ERPs of all test statistics are very small and in the vicinity of the nominal size or even below when $\phi = .9$. For all designs, ERPs under the null are typically nonincreasing functions of S_n for all tests in the study and in most

[10]The only exception under the null is the test based on K_{EL,H_2} that has higher ERPs for intermediate bandwidth values for $\rho_{uV} = .5$ for the case $k = 20$ and $\Pi_1 = .01$ across almost all AR/MA designs.

[11]Two exceptions to this statement about power for $\Pi_1 = .01$ are (1) the case of same AR parameter $\phi = .9$, where the power of all the test statistics over all parameter combinations is higher than 50% for small numbers of S_n and (2) all other MA/AR cases for $k = 20$ and $\rho_{uV} = 0$, where the power of K_{EL,H_2} can reach values of up to 40% for some intermediate bandwidths.

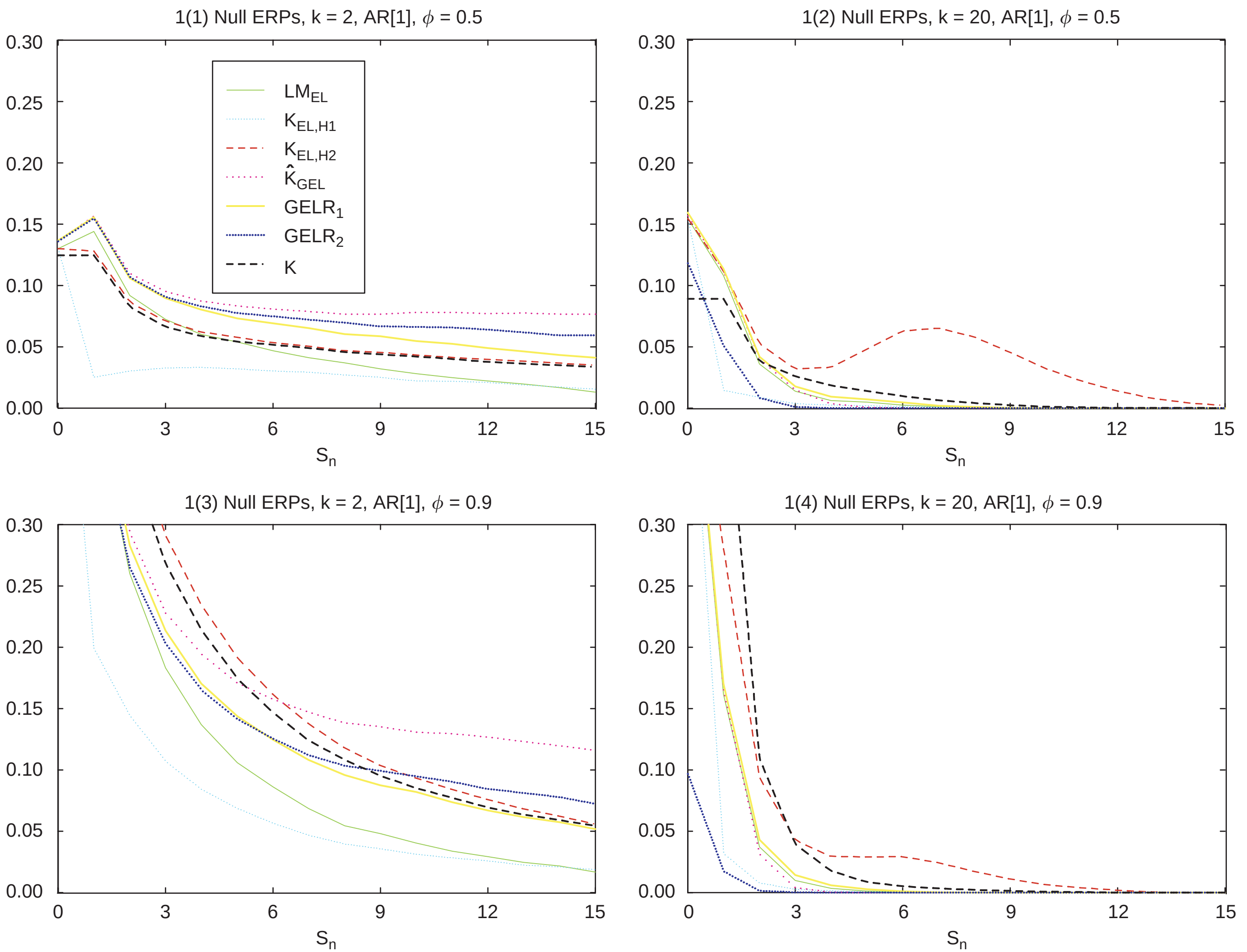
1(1) Null ERPs, k = 2, AR[1], ϕ = 0.5
1(2) Null ERPs, k = 20, AR[1], ϕ = 0.5
1(3) Null ERPs, k = 2, AR[1], ϕ = 0.9
1(4) Null ERPs, k = 20, AR[1], ϕ = 0.9
LM_{EL}
$K_{EL,H1}$
$K_{EL,H2}$
$\hat{K}_{GEL}$
$GELR_1$
$GELR_2$
K
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
3
6
9
12
15
S_n

Fig. 1.

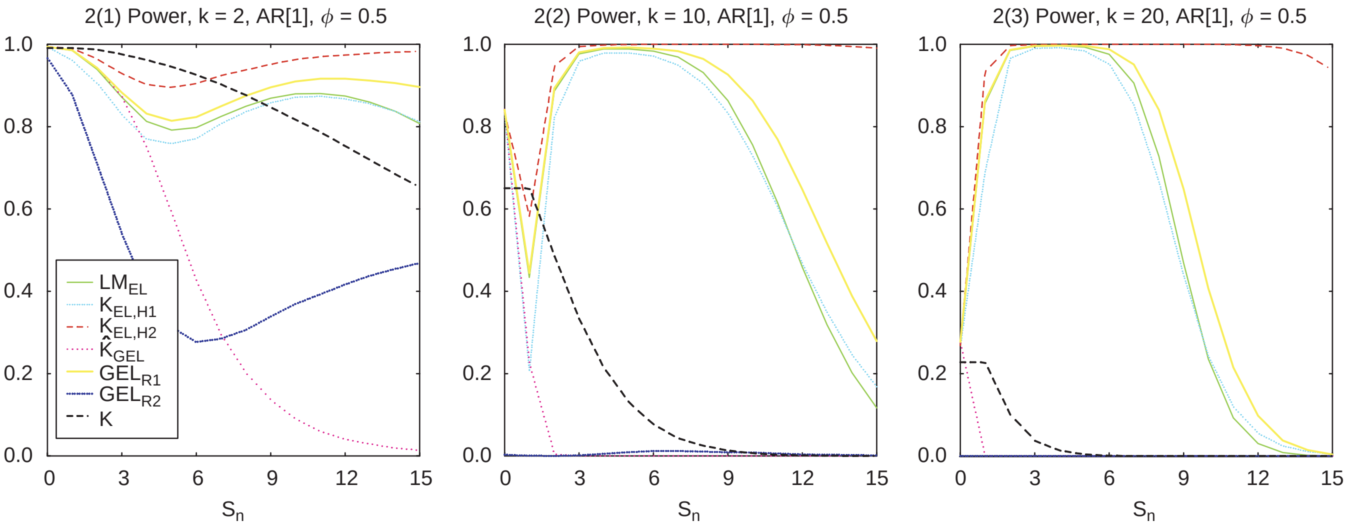
2(1) Power, k = 2, AR[1], ϕ = 0.5
2(2) Power, k = 10, AR[1], ϕ = 0.5
2(3) Power, k = 20, AR[1], ϕ = 0.5
1.0
0.8
0.6
0.4
0.2
0.0
0
3
6
9
12
15
S_n
LM_{EL}
$K_{EL,H1}$
$K_{EL,H2}$
$\hat{K}_{GEL}$
GEL_{R1}
GEL_{R2}
K

Fig. 2.

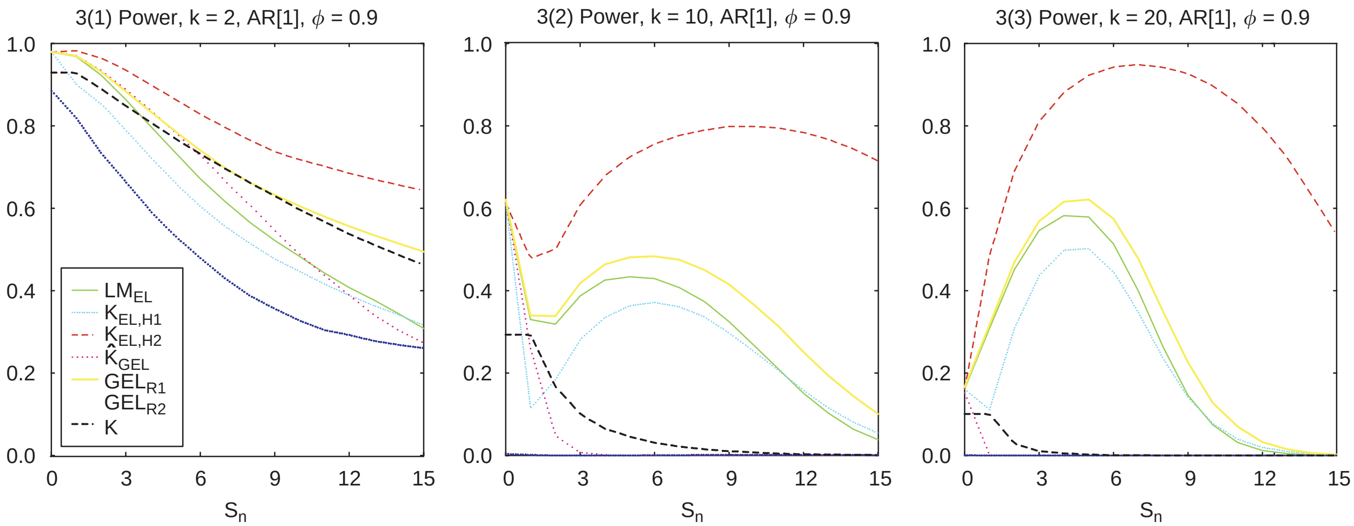
3(1) Power, k = 2, AR[1], ϕ = 0.9
3(2) Power, k = 10, AR[1], ϕ = 0.9
3(3) Power, k = 20, AR[1], ϕ = 0.9
LM_{EL}
$K_{EL,H1}$
$K_{EL,H2}$
$\hat{K}_{GEL}$
GEL_{R1}
GEL_{R2}
K
1.0
0.8
0.6
0.4
0.2
0.0
0
3
6
9
12
15
S_n

Fig. 3.

cases the maximum smoothing number $S_n = 15$ considered here is enough to reduce ERPs to about the nominal level or even less. However, for various scenarios with few instruments, Otsu's (2006) $\widehat{K}_{\mathrm{GEL}}$ test continues to overreject even for $S_n = 15$, see Figs. 1(1) and especially 1(3), where $k = 2$. The computationally simpler modifications of $\widehat{K}_{\mathrm{GEL}}$ from Theorem 3 improve on the size distortion of $\widehat{K}_{\mathrm{GEL}}$, see Figs. 1(1 and 3). Across virtually all designs and uniformly in $S_n \geqslant 1$, the ERPs under the null of the test statistic K_{EL,H_1} are smallest among all test statistics considered and with the exception of few highly persistent designs (such as in Fig. 1(3) where $\phi = .9$) the ERPs of this test equal or are below the nominal size for any $S_n \geqslant 1$. The two closely related statistics LM_{EL} and K_{EL,H_2} typically require more smoothing to reduce the ERPs under the null below the nominal size and in few highly persistent cases (such as in Fig. 1(3) where $\phi = .9$) $S_n = 15$ is not even quite sufficient to control size for K_{EL,H_2}. Comparing K_{EL,H_1} on the one and LM_{EL} and K_{EL,H_2} on the other side, the former statistic oftentimes leads to a quite conservative test which has negative effects on power relative to the other statistics as seen below. In that respect, LM_{EL} seems to offer a good compromise between the two hybrid statistics in terms of size and power trade-off. Recalling the construction of the hybrid statistics in (2.31) and (2.32), one might expect the performance of the hybrid tests to be in between the ones of the LM_{EL} and K test, with K_{EL,H_1} and K_{EL,H_2} being closer to LM_{EL} and K, respectively. The Monte Carlos do not confirm this expectation. While K_{EL,H_1} is typically smallest, there is no simple ranking among LM_{EL}, K_{EL,H_2}, and K; e.g. compare Figs. 1(2/4), where the ERPs of the K test for small S_n are far smaller/higher than for the K_{EL,H_2} test. While replacing $\widehat{g}_n(\theta)$ in LM_ρ by $2\widehat{g}(\theta)$ in $K_{\rho,H_1}(\theta)$ uniformly decreases ERPs, this effect is oftentimes overcompensated by replacing $\widehat{\Delta}(\theta)$ in $K_{\rho,H_1}(\theta)$ by $2\widehat{\Delta}(\theta)$ in $K_{\rho,H_2}(\theta)$. The latter statistic differs from K only through the matrix D_ρ but no consistent ranking in terms of size of the two tests can be derived from our simulation study.

Summarizing we find that for sufficient smoothing, the testing procedures have ERPs under the null that come close to the nominal size. One exception is the test based on the statistic $\widehat{K}_{\mathrm{GEL}}$ that seems to somewhat overreject even for $S_n = 15$ when k is small.

Next the *power* results are summarized. We first discuss the separate effects of k, ϕ, and S_n on the power properties of the tests. It seems that increasing k has a negative impact on the power properties of $\widehat{K}_{\mathrm{GEL}}$ and K (see Figs. 2(1– 3) and 3(1– 3)). On the other hand, for LM_{EL} and K_{ρ,H_j}, for $j = 1, 2$, the effect of k on power is mixed and seems to depend on the bandwidth S_n. For example, in Figs. 2(1 and 2), power decreases for increasing k for small and large bandwidths S_n but increases for increasing k for intermediate bandwidths S_n. Increasing the AR coefficient ϕ generally seems to have a negative impact on power (compare Figs. 2 and 3). While the power of $\widehat{K}_{\mathrm{GEL}}$ and K seems to be a decreasing function of S_n, the effect of the bandwidth on the power of the other statistics depends on the scenario. For example, in Fig. 3(1) power decreases in S_n while in all the other figures power is not a monotonic function of S_n.

Next we compare the power properties of the tests to each other. Overall, K_{ρ,H_2} seems to have best power properties across all statistics considered. The power gains over the other tests can be dramatic in cases of large k and high ϕ, see Fig. 3. When $k = 2$ and $\phi = .5$, the K test takes on the power lead for small values of the bandwidth, see Fig. 2(1). However, the power of K and especially the power of $\widehat{K}_{\mathrm{GEL}}$ is very low relative to the other tests when k is large; even when $k = 2$ the power loss can be dramatic when S_n is large, see Fig. 2(1). With regards to power there seems to be a consistent ranking of LM_{EL}, K_{ρ,H_1}, and K_{ρ,H_2} with LM_{EL} having power between K_{ρ,H_1} and K_{ρ,H_2}. In that respect, LM_{EL} seems to offer a good trade-off between the excellent size and power properties of K_{ρ,H_1} and K_{ρ,H_2}, respectively. Given the sometimes large differences in power between K_{ρ,H_2} and K, we conclude that the components D_ρ and D_θ in these statistics have an important impact on the performance of the tests. With respect to the statistics $GELR_{\mathrm{ET}}(\theta_0, \mu_{\mathrm{EL}}(\theta_0))$ and $GELR_{\mathrm{ET}}(\theta_0, \tilde{\mu}_{\mathrm{EL}}(\theta_0))$ we find that overall the former has very competitive while the latter has very poor power properties.

GS found that the comparative advantage of GEL-based tests in i.i.d. simulations occurs in situations with thick tailed or asymmetric error distributions. Here, we find that even with normal errors, GEL-based tests can outperform the K test, depending on the scenario, most crucially the number of instruments.

In summary we find that both the finite–sample size and power properties of the tests based on the new statistic LM_{EL} are very competitive. The new hybrid tests K_{ρ,H_1} and K_{ρ,H_2} provide very good size and power properties, respectively. Based on our simulations, we also recommend the statistic $GELR_{\mathrm{ET}}(\theta_0, \mu_{\mathrm{EL}}(\theta_0))$.

Acknowledgments

We would like to thank the Co-Editor, P. Robinson, an anonymous Associate Editor, and two referees for very helpful comments. This paper is an extension of the first chapter of Guggenberger's dissertation at Yale University and was written in August 2003. Guggenberger thanks his advisor D.W.K. Andrews who played a very important role in the writing process. We would also like to thank J. Hahn and T. Vogelsang for detailed comments and seminar participants at Brown, Harvard/MIT, Maryland, NYU, and SUNY Albany in 2003, at Irvine, Laval, Riverside, and San Diego in 2004, at the Second Conference on Information and Entropy Econometrics in honor of George Judge in Washington in 2005, and at the IWH Halle and Bonn in 2006. Guggenberger and Smith gratefully acknowledge research support from a faculty research grant from UCLA in 2005 and financial support through a 2002 Leverhulme Major Research Fellowship, respectively.

Appendix

Additional notation is given and then the assumptions for Theorem 1 are stated.

As discussed above, for the validity of the tests in Theorem 1, consistency of $\widehat{\Delta}(\theta_0)/2$ in (2.6) for the long-run variance matrix $\Delta(\theta_0)$ is essential. To show consistency of $\widehat{\Delta}(\theta_0)/2$, we assume consistency of the classical Bartlett kernel HAC estimator (which holds under appropriate assumptions given in Andrews, 1991, Proposition 1) and then show that the HAC estimator differs from $\widehat{\Delta}(\theta_0)/2$ by a $o_p(1)$ term only. The latter is similar to Lemmas 2.1 and A.3 in Smith (2001, 2005). The same procedure can be applied to other long-run variance expressions, such as $\Delta_A(\theta_0)$, defined in M_{θ_0}(vii) below and its corresponding estimator $\widehat{\Delta}_A(\theta_0)/2$, where

$$\widehat{\Delta}_A(\theta_0) := S_n \sum_{i=1}^{n} (vec\, G_{inA}(\theta_0)) g'_{in}(\theta_0)/n. \tag{A.1}$$

We now give the details.

In (2.17), decompose $G_i(\theta)$ into $(G_{iA}(\theta), G_{iB}(\theta))$, where $G_{iA}(\theta) \in \mathbb{R}^{k\times p_A}$ and $G_{iB}(\theta) \in \mathbb{R}^{k\times p_B}$.

Denote by k^* the Bartlett kernel given by

$$k^*(x) := 1 - |x/2| \quad \text{if } |x| \leqslant 2 \quad \text{and} \quad k^*(x) = 0 \quad \text{otherwise}. \tag{A.2}$$

The Bartlett kernel is essentially the convolution of the truncated kernel, in fact, $k^*(x) = \int k(x-y)k(y)\,dy/2$, see Smith (2001, Example 2.1). The Bartlett HAC estimator of the long–run covariance between sequences of mean zero random vectors $r = (r_i)_{i=1,\dots,n}$ and $s = (s_i)_{i=1,\dots,n}$ is given by

$$\tilde{J}_n(r,s) := \sum_{j=-n+1}^{n-1} k^*(j/S_n)\tilde{\Gamma}_j(r,s), \quad \text{where}$$
$$\tilde{\Gamma}_j(r,s) := \begin{cases} \sum_{i=j+1}^{n} r_i s'_{i-j}/n & \text{for } j \geqslant 0, \\ \sum_{i=-j+1}^{n} r_{i+j} s'_i/n & \text{for } j < 0, \end{cases} \tag{A.3}$$

see Andrews (1991, eq. (3.2)). Under certain assumptions, that include stationarity, it can be shown that (see Andrews, 1991, Assumption A, Proposition 1)

$$\tilde{J}_n(g_i, g_i) \to_p \Delta, \quad \tilde{J}_n(vec\, G_{iA}, g_i) \to_p \Delta_A, \tag{A.4}$$

where the argument θ_0 was left out to simplify notation. Below it is shown that the Bartlett HAC estimator and $\widehat{\Delta}(\theta_0)/2$ have the same probability limit.[12] Therefore, assuming (A.4) and some technicalities, $\widehat{\Delta}(\theta_0)/2$ is consistent for the long–run variance $\Delta(\theta_0)$. The same statement is true for $\Delta_A(\theta_0)$ and its estimator.

[12]Note that the assumptions $\tilde{J}_n(vec\, G_{iA}, g_i) \to_p \Delta_A$ and $\tilde{J}_n(vec(G_{iA} - \mathrm{E}G_{iA}), g_i) \to_p \Delta_A$ are equivalent under weak conditions, for example, under stationarity. Therefore, for consistency of the HAC estimator the possibly non-zero mean of $vec\, G_{iA}$ does not matter as long as

A.1. Assumptions

A.1.1. Full-vector tests

The assumptions of Theorem 1 are now stated and discussed. For the asymptotic distribution of $GELR_\rho$ Assumption M_{θ_0}(i)–(iii) are needed. For the statistics $LM_\rho(\theta_0)$ and $S_\rho(\theta_0)$ we also need M_{θ_0}(iv)–(vii). For notational simplicity, the argument θ_0 is left out in M_{θ_0}(v)–(vii) and in the following discussion. Denote by Z the set of integer numbers.

Assumption M_{θ_0}. Suppose (i) $\max_{1\leqslant i\leqslant n}\|g_i(\theta_0)\| = o_p(S_n^{-1}n^{1/2})$; (ii) for $S_n\to\infty$ and $S_n = o(n^{1/2})$ we have $\widetilde{J}_n((g_i(\theta_0)),(g_i(\theta_0)))\to_p \Delta(\theta_0)>0$; $\sup_{i,j\geqslant 1}\mathrm{E}\|g_i(\theta_0)g_j'(\theta_0)\|<\infty$; for any sequence $m\to\infty$ and $m=o(n^{1/2})$, $\sup_{k\in Z}\mathrm{E}\|(1/nm)\sum_{j=1}^n\sum_{i=k}^{k+m}g_{j+i}(\theta_0)g_j'(\theta_0)\| = o(1)$; $S_n n^{-1}\sum_{i=1}^n\|g_{in}(\theta_0)g_{in}(\theta_0)'\| = O_p(1)$; (iii) $\Psi_n(\theta_0)\to_d\Psi(\theta_0)$, where $\Psi(\theta_0)\equiv \mathrm{N}(0,\Delta(\theta_0))$;

$$\text{(iv) } M_{1n}(\theta_0):=(\partial m_{1n}/\partial\theta)|_{\theta=\theta_0}\to M_1(\theta_0):=(\partial m_1/\partial\theta)|_{\theta=\theta_0}\in\mathbb{R}^{k\times p}, \tag{A.5}$$

$$\mathrm{E}\widehat{G}(\theta_0) = n^{-1/2}M_{1n}(\theta_0)+(0,M_2(\beta_0))\to(0,M_2(\beta_0)); \tag{A.6}$$

(v) $\widetilde{J}_n((vec\, G_{iA}),(g_i))\to_p\Delta_A$ (Δ_A is defined in (vii)); $\sup_{i,j\geqslant 1}\mathrm{E}\|vec\, G_{iA}g_j'\|<\infty$; for any sequence $m\to\infty$ and $m=o(n^{1/2})$, $\sup_{k\in Z}\mathrm{E}\|(1/nm)\ \sum_{j=1}^n\sum_{i=k}^{k+m}vec\, G_{j+iA}g_j'\| = o(1)$; $\widehat{G}_B\to_p\mathrm{E}\widehat{G}_B$; (vi) $\max_{1\leqslant i\leqslant n}\|G_{iA}\| = o_p(S_n^{-1}n^{1/2})$; $S_n n^{-1}\sum_{i=1}^n\|vec\, G_{inA}g_{in}'\| = O_p(1)$; $\max_{1\leqslant i\leqslant n}\|G_{iB}\| = o_p(S_n^{-1}n)$; $S_n n^{-3/2}\sum_{i=1}^n\|vec\, G_{inB}g_{in}'\| = o_p(1)$; (vii) $n^{-1/2}\sum_{i=1}^n((vec(G_{iA}-EG_{iA}))',g_i')'\to_d\mathrm{N}(0,V)$, where

$$V:=\lim_{n\to\infty}var\left(n^{-1/2}\sum_{i=1}^n(vec\, G_{iA}',g_i')\right)'\in\mathbb{R}^{k(p_A+1)\times k(p_A+1)} \tag{A.7}$$

has full column rank. Decompose V into

$$V=\begin{pmatrix}\Delta_{AA} & \Delta_A\\ \Delta_A' & \Delta\end{pmatrix},\quad\text{where } \Delta_{AA}\in\mathbb{R}^{p_Ak\times p_Ak}. \tag{A.8}$$

A discussion of Assumption M_{θ_0} now follows. Assuming $S_n = cn^\alpha$ for positive constants c and $\alpha<\frac{1}{2}$, a sufficient condition for M_{θ_0}(i) is given by the moment condition $\sup_{i\geqslant 1}\mathrm{E}\|g_i(\theta_0)\|^\xi<\infty$ for some $\xi>2/(1-2\alpha)$, see GS, Eq. (2.4), for a similar statement and a proof. Analogous sufficient conditions can be formulated for M_{θ_0}(vi).

The high-level assumption $\widetilde{J}_n((g_i),(g_i))\to_p\Delta$ in M_{θ_0}(ii) is satisfied under sufficient conditions given in Andrews (1991, Proposition 1) which include stationarity. We prefer the high-level assumption to the sufficient condition because it may hold even when the data are not stationary, e.g. in cases of non-identically distributed data. M_{θ_0}(ii) then guarantees that $\widehat{\Delta}\to_p 2\Delta$, see Lemma 2 below. The technical assumption $\sup_{k\in Z}\mathrm{E}\|(1/nm)\sum_{j=1}^n\sum_{i=k}^{k+m}g_{j+i}g_j'\| = o(1)$ can be interpreted as a mild form of mixing, see also analogous assumptions in M_{θ_0}(v), and is needed in the proof of Lemma 2. The assumption $S_n n^{-1}\sum_{i=1}^n\|g_{in}g_{in}'\| = O_p(1)$ is needed in the proof of Theorem 1(i) to show that $S_n\sum_{i=1}^n(\rho_2(\widetilde{\lambda}'g_{in})+1)g_{in}g_{in}'/n$ is $o_p(1)$. To motivate this assumption, note that if a CLT holds then we have $g_{in} = O_p(S_n^{-1/2})$. The analogous assumptions in M_{θ_0}(v) and (vi) are needed in deriving (A.28) and can be motivated in the same manner, noting that also $vec\, G_{inA} = O_p(S_n^{-1/2})$ by M_{θ_0}(iv) and (vii).

M_{θ_0}(iii) is the "high-level" assumption also used in Stock and Wright (2000).

(*footnote continued*)
$\mathrm{E}g_i = 0$. More precisely, it can be shown that under stationarity

$$\widetilde{J}_n(vec\, G_{iA},g_i)-\widetilde{J}_n(vec(G_{iA}-\mathrm{E}G_{iA}),g_i) = \widetilde{J}_n(vec\,\mathrm{E}G_{iA},g_i)\to_p 0.$$

This can be shown by establishing that for any $s=1,\dots,p_Ak$ and $t=1,\dots,k$ and for some $c<\infty$ it holds that $(\mathrm{E}\widetilde{J}_n(vec\,\mathrm{E}G_{iA},g_i))_{s,t}=0$ and $(n/S_n^2)\mathrm{E}(\widetilde{J}_n(vec\,\mathrm{E}G_{iA},g_i))_{s,t}^2\leqslant c$, see Hannan (1970, p. 280) for similar calculations. Because by assumption $(n/S_n^2)\to\infty$, the latter implies consistency.

A sufficient condition for M_{θ_0}(iv) is given by: for some open neighborhood $\mathscr{M} \subset \Theta$ of θ_0, $\widehat{g}(\cdot)$ is differentiable at $\overline{\theta}$ a.s. for each $\overline{\theta} \in \mathscr{M}$, $\widehat{g}(\overline{\theta})$ is integrable for all $\overline{\theta} \in \mathscr{M}$ (with respect to the probability measure), $\sup_{\overline{\theta} \in \mathscr{M}} \|\widehat{G}(\overline{\theta})\|$ is integrable, $m_{1n} \in C^1(\Theta)$, and $M_{1n}(\cdot)$ converges uniformly on Θ to some function. These conditions allow the interchange of the order of integration and differentiation in Assumption ID, i.e. $(\partial \mathrm{E}\widehat{g}/\partial\theta)|_{\theta=\theta_0} = \mathrm{E}\widehat{G}(\theta_0)$. Note that by ID the limit matrix $(0, M_2(\beta_0))$ is singular of rank p_B.

Let

$$\widehat{G}_n(\theta) := n^{-1} \sum_{i=1}^{n} G_{in}(\theta) \tag{A.9}$$

and decompose $\widehat{G}_n(\theta)$ as $(\widehat{G}_{nA}(\theta), \widehat{G}_{nB}(\theta))$, where $\widehat{G}_{nA}(\theta) \in \mathbb{R}^{k \times p_A}$ and $\widehat{G}_{nB}(\theta) \in \mathbb{R}^{k \times p_B}$. The assumption $\max_{1 \leqslant i \leqslant n} \|G_{iB}\| = o_p(S_n^{-1} n)$ in M_{θ_0}(vi) ensures that $\widehat{G}_{nB} - 2\widehat{G}_B = o_p(1)$. This can be shown along the lines of Lemma 1.

Besides technical assumptions, M_{θ_0} essentially states that the HAC estimator $\widetilde{J}_n$ is consistent (parts (ii) and (v)) and that a CLT holds for $((vec(G_{iA} - EG_{iA}))', g_i')'$ (parts (iii) and (vii)). For the latter, primitive sufficient conditions based on mixing properties can be stated along the lines of Wooldridge and White (1988). The CLT assumption is very closely related to Assumption 1 in Kleibergen (2005a). Assumption (v) needs to be substituted by an assumption analogous to (2.19) when dealing with the unsmoothed statistics. When deriving the limit distribution of $S_\rho^*(\theta_0)$ and $LM_\rho^*(\theta_0)$ we assume

$$\sum_{i=1}^{n} vec\, G_{iA}(\theta_0) g_i'(\theta_0)/n \rightarrow_p \Omega_A(\theta_0) := \lim_{n\to\infty} \mathrm{E} \sum_{i=1}^{n} vec\, G_{iA}(\theta_0) g_i'(\theta_0)/n. \tag{A.10}$$

A.1.2. Sub-vector tests

For the sub-vector tests we give high-level assumptions. More primitive assumptions along the lines of Assumption M_{θ_0} could be stated at the cost of additional space.

Let $\widehat{G}_{A_j}(\theta) := n^{-1} \sum_{i=1}^{n} (\partial g_i/\partial \alpha_j)(\theta)$ and likewise $\widehat{G}_{nA_j}(\theta) := n^{-1} \sum_{i=1}^{n} (\partial g_{in}/\partial \alpha_j)(\theta)$.

Assumption M_{α_0}. For any consistent estimators $\widetilde{\beta}, \overline{\beta} \rightarrow_p \beta_0$ we have (i) $\max_{1 \leqslant i \leqslant n} \sup_{\beta \in B} \|g_i(\theta_\beta)\| = o_p(S_n^{-1} n^{1/2})$; $S_n^{-1} n^{-1} \sum_{i=1}^{n} g_i(\widehat{\theta}_0) = o_p(1)$; (ii) for $S_n \to \infty$, $S_n = o(n^{1/2})$ we have $\widehat{\Delta}(\theta_{\widehat{\beta}}) \rightarrow_p 2\Delta(\theta_0) > 0$; $\lambda_{\max}(\widehat{\Delta}(\widehat{\theta}_0))$ is bounded w.p.a.1; $S_n n^{-1} \sum_{i=1}^{n} \|g_{in}(\theta_{\widehat{\beta}}) g_{in}(\theta_{\widehat{\beta}})'\| = O_p(1)$; (iii) $\widehat{G}_B(\theta_{\widehat{\beta}})$ exists; $\widehat{G}_B(\theta_{\widehat{\beta}}) \rightarrow_p \mathrm{E}\widehat{G}_B(\theta_{\widehat{\beta}}) = (n^{-1/2} (\partial m_{1n}/\partial\beta)(\theta_{\widehat{\beta}}) + (\partial m_2/\partial\beta)(\alpha_{02}, \widetilde{\beta}) \rightarrow (\partial m_2/\partial\beta)(\alpha_{02}, \beta_0)$; $n^{-1} S_n^{-1} \sum_{i=1}^{n} G_{iB}(\theta_{\widehat{\beta}}) = o_p(1)$; $\max_{1 \leqslant i \leqslant n} \|G_{iB}(\theta_{\widehat{\beta}})\| = o_p(S_n^{-1} n)$; (iv) $\widehat{g}(\widehat{\theta}_0) \rightarrow_p \mathrm{E}\widehat{g}(\widehat{\theta}_0)$; $\Psi_n(\theta_0) \rightarrow_d \Psi(\theta_0)$, where $\Psi(\theta_0) \equiv \mathrm{N}(0, \Delta(\theta_0))$; (v) $(\partial\, vec\, \widehat{G}_{A_1}/\partial\beta)(\theta)$ exists on a neighborhood of θ_0 and $(\partial\, vec\, \widehat{G}_{A_1}/\partial\beta)(\theta_{\widehat{\beta}}) \rightarrow_p 0$; (vi) $\max_{1 \leqslant i \leqslant n} \|G_{iA_1}(\theta_{\widehat{\beta}})\| = o_p(S_n^{-1} n^{1/2})$; $n^{-1/2} \sum_{i=1}^{n} ((vec(G_{iA_1}(\theta_0) - \mathrm{E}G_{iA_1}(\theta_0)))', (g_i(\theta_0) - \mathrm{E}g_i(\theta_0))')' \rightarrow_d \mathrm{N}(0, V^\alpha)$, where V^α is the appropriate submatrix of V defined in M_{θ_0}(vii); $V^\alpha > 0$; (vii) $S_n n^{-1} \sum_{i=1}^{n} vec(G_{inA_1}(\theta_{\widehat{\beta}})) g_{in}(\theta_{\widehat{\beta}})' \rightarrow_p 2\Delta_{A_1}$ (defined in (A.11)); $S_n n^{-1} \sum_{i=1}^{n} \|vec\, G_{inA_1}(\theta_{\widehat{\beta}}) g_{in}(\theta_{\widehat{\beta}})'\| = O_p(1)$; $S_n n^{-3/2} \sum_{i=1}^{n} \|vec\, G_{inA_2}(\theta_{\widehat{\beta}}) g_{in}(\theta_{\widehat{\beta}})'\| = o_p(1)$; similar to (iii), $\widehat{G}_{A_2}(\theta_{\widehat{\beta}})$ exists and $\widehat{G}_{A_2}(\theta_{\widehat{\beta}}) \rightarrow_p (\partial m_2/\partial\beta)\ (\alpha_{02}, \beta_0)$; $\max_{1 \leqslant i \leqslant n} \|G_{iA_2}(\theta_{\widehat{\beta}})\| = o_p(S_n^{-1} n)$.

In M_{α_0}(vi) write

$$V^\alpha = \begin{pmatrix} \Delta_{A_1 A_1} & \Delta_{A_1} \\ \Delta_{A_1}' & \Delta \end{pmatrix} \quad \text{where } \Delta_{A_1 A_1} \in R^{p_{A_1} k \times p_{A_1} k}. \tag{A.11}$$

Mutatis mutandis the assumptions in M_{α_0} can be interpreted as their counterparts in M_{θ_0}. For example, M_{α_0}(ii) guarantees that $\lambda_{\min}(\widehat{\Delta}(\widehat{\theta}_0))$ is bounded away from zero w.p.a.1. which is needed when deriving a slight variation of Lemma 4. Sufficient conditions for the high-level assumptions above can be given along the lines of GS, e.g. for $(\partial\, vec\, \widehat{G}_{A_1}/\partial\beta)(\theta_{\widehat{\beta}}) \rightarrow_p 0$ in M_{α_0}(v), see their M_α(v), (vii), and ID_α. Likewise, sufficient conditions stated in terms of HAC estimators can be given for M_{α_0}(ii) and the first part of M_{α_0}(vii); see also M_α(viii) in GS for more primitive conditions.

A.2. Proofs

The next lemmas are helpful in the proof of the main result. Note that the assumptions made in Lemma 1 are implied by M_{θ_0}(i), (iii), (vi), and (vii), e.g. $\widehat{G}_A(\theta_0) = O_p(n^{-1/2})$ follows from M_{θ_0}(vii) and Eq. (A.5). Recall $\widehat{G}_{nA}(\theta) = n^{-1}\sum_{i=1}^n G_{inA}(\theta)$.

Lemma 1. *Suppose $S_n \to \infty$ and $S_n = o(n^{1/2})$.*

$$\text{If } \max_{1\leqslant i\leqslant n} \|g_i\| = o_p(S_n^{-1}n^{1/2}), \widehat{g} = O_p(n^{-1/2}) \quad \text{then } n^{1/2}(\widehat{g}_n - 2\widehat{g}) = o_p(1).$$

$$\text{If } \max_{1\leqslant i\leqslant n} \|G_{iA}\| = o_p(S_n^{-1}n^{1/2}), \widehat{G}_A = O_p(n^{-1/2}) \quad \text{then } n^{1/2}(\widehat{G}_{nA} - 2\widehat{G}_A) = o_p(1),$$

where again θ_0 is left out to simplify the notation.

Proof. For the first equation tedious but straightforward calculations imply that

$$\begin{aligned}
n^{-1}\sum_{i=1}^n g_{in} &= n^{-1}\sum_{i=1}^n S_n^{-1}\sum_{j=i-n}^{i-1} k(j/S_n)g_{i-j} = n^{-1}\sum_{i=1}^n S_n^{-1}\sum_{j=\max(i-n,-S_n)}^{\min(i-1,S_n)} g_{i-j} \\
&= n^{-1}\sum_{i=S_n+1}^{n-S_n} \frac{2S_n+1}{S_n} g_i + n^{-1}\sum_{i=1}^{S_n} \frac{S_n+i}{S_n} g_i + n^{-1}\sum_{i=n-S_n+1}^{n} \frac{n-i+S_n+1}{S_n} g_i \\
&= 2n^{-1}\sum_{i=1}^n g_i + n^{-1}\sum_{i=S_n+1}^{n-S_n} \frac{1}{S_n} g_i \\
&\quad + n^{-1}\sum_{i=1}^{S_n} \frac{i-S_n}{S_n} g_i + n^{-1}\sum_{i=n-S_n+1}^{n} \frac{-S_n+n-i+1}{S_n} g_i \\
&= 2n^{-1}\sum_{i=1}^n g_i + o_p(n^{-1/2}),
\end{aligned} \tag{A.12}$$

where the last equation uses $\max_{1\leqslant i\leqslant n}\|g_i\| = o_p(S_n^{-1}n^{1/2})$ and $\widehat{g} = O_p(n^{-1/2})$ to show that the remainder terms are $o_p(n^{-1/2})$. The proof of the second equation can be derived in exactly the same way. □

It is now shown that under M_{θ_0}, $\widehat{\Delta}/2$ and $\widehat{\Delta}_A/2$ are consistent for Δ and Δ_A. The first part of the following lemma is similar to Lemma A.3 in Smith (2001). Note that the assumptions in the lemma are part of M_{θ_0}(ii) and (v).

Lemma 2. *For $S_n \to \infty$ assume $S_n = o(n^{1/2})$. If $\sup_{i,j\geqslant 1} E\|g_i g_j'\| < \infty$ and $\sup_{k\in Z} E\|(1/nS_n)\sum_{j=1}^n \sum_{i=k}^{k+S_n} g_{j+i}g_j'\| = o(1)$ then*

$$\widehat{\Delta} - 2\widetilde{J}_n((g_i),(g_i)) = o_p(1).$$

If $\sup_{i,j\geqslant 1} E\|vec\, G_{iA}g_j'\| < \infty$ and $\sup_{k\in Z} E\|(1/nS_n)\sum_{j=1}^n \sum_{i=k}^{k+S_n} vec\, G_{j+iA}g_j'\| = o(1)$ then

$$\widehat{\Delta}_A - 2\widetilde{J}_n((vec\, G_{iA}),(g_i)) = o_p(1), \tag{A.13}$$

where the argument θ_0 is left out to simplify the notation.

Proof. For the first statement easy calculations lead to

$$2\widetilde{J}_n((g_i),(g_i))-\widehat{\Delta}=\sum_{i=-n+1}^{n-1}n^{-1}\sum_{j=\max(1,1-i)}^{\min(n,n-i)}k_{ij}g_{j+i}g_j' \quad \text{for}$$

$$k_{ij}:=2k^*(i/S_n)-S_n^{-1}\sum_{l=1-j}^{n-j}k((l-i)/S_n)k(l/S_n). \tag{A.14}$$

Using the definitions of k and k^* tedious calculations show that for $0\leqslant i<S_n$

$$k_{ij}=\begin{cases} S_n^{-1}(S_n-i-j) & \text{for } 1\leqslant j\leqslant S_n-i+1,\\ -S_n^{-1} & \text{for } S_n-i+1<j\leqslant n-S_n,\\ -S_n^{-1}(n-j-S_n+1) & \text{for } n-S_n<j\leqslant n-i,\end{cases} \tag{A.15}$$

that for $-S_n<i<0$

$$k_{ij}=\begin{cases} S_n^{-1}(S_n-j) & \text{for } 1-i\leqslant j\leqslant S_n+1,\\ -S_n^{-1} & \text{for } S_n+1<j<n-S_n-i,\\ S_n^{-1}(S_n+i-n+j-1) & \text{for } n-S_n-i\leqslant j\leqslant n,\end{cases} \tag{A.16}$$

that $k_{ij}=-S_n^{-1}$ if $S_n\leqslant|i|\leqslant 2S_n$ and that $k_{ij}=0$ otherwise. Using the moment assumptions, it then follows that $2\widetilde{J}_n((g_i),(g_i))-\widehat{\Delta}$ reduces to $o_p(1)$ expressions. For example, by Markov's inequality the summand $\Pr(\|-\sum_{i=S_n}^{2S_n}n^{-1}S_n^{-1}\sum_{j=1}^{n-i}g_{j+i}g_j'\|>\varepsilon)$ can be bounded by

$$\varepsilon^{-1}S_n^{-1}n^{-1}\,\mathrm{E}\left\|\sum_{i=S_n}^{2S_n}\sum_{j=1}^{n-i}g_{j+i}g_j'\right\|=\varepsilon^{-1}S_n^{-1}n^{-1}\,\mathrm{E}\left\|\sum_{j=1}^{n-S_n}\sum_{i=S_n}^{\min(n-j,2S_n)}g_{j+i}g_j'\right\|. \tag{A.17}$$

Using $\sup_{k\in Z}\mathrm{E}\|(1/nS_n)\sum_{j=1}^{n}\sum_{i=k}^{k+S_n}g_{j+i}g_j'\|=o(1)$ it then follows that the RHS of this expression is o(1). The proof of the second claim is completely analogous and therefore omitted. □

Given the results in Lemma 1 and consistency of $\widehat{\Delta}/2$ and $\widehat{\Delta}_A/2$, the proof of Theorem 1 is along the same lines as the proofs of Theorems 3 and 4 in GS.

As in GS, the proof hinges on the following two lemmas. Let $c_n:=S_nn^{-1/2}\max_{1\leqslant i\leqslant n}\|g_{in}(\theta_0)\|$. Let $\Lambda_n:=\{\lambda\in\mathbb{R}^k:\|\lambda\|\leqslant S_nn^{-1/2}c_n^{-1/2}\}$ if $c_n\neq 0$ and $\Lambda_n=\mathbb{R}^k$ otherwise.

Lemma 3. *Assume* $\max_{1\leqslant i\leqslant n}\|g_i(\theta_0)\|=o_p(S_n^{-1}n^{1/2})$. *Then* $\sup_{\lambda\in\Lambda_n,1\leqslant i\leqslant n}|\lambda' g_{in}(\theta_0)|\to_p 0$ *and* $\Lambda_n\subset\widehat{\Lambda}_n(\theta_0)$ *w.p.a.*1.

Proof. The case $c_n=0$ is trivial and thus w.l.o.g. $c_n\neq 0$ can be assumed. Note that $\|g_{in}(\theta_0)\|\leqslant S_n^{-1}\sum_{j=i-n}^{i-1}k(j/S_n)\|g_{i-j}(\theta_0)\|$ and thus by the definition of $k(\cdot)$

$$\begin{aligned}\max_{1\leqslant i\leqslant n}\|g_{in}(\theta_0)\|&\leqslant\max_{1\leqslant i\leqslant n}S_n^{-1}\sum_{j=\max(-S_n,i-n)}^{\min(S_n,i-1)}\|g_{i-j}(\theta_0)\|\\ &\leqslant(2S_n+1)S_n^{-1}\max_{1\leqslant i\leqslant n}\|g_i(\theta_0)\|=o_p(S_n^{-1}n^{1/2}).\end{aligned} \tag{A.18}$$

Therefore, $c_n=o_p(1)$ and the first part of the statement follows from

$$\begin{aligned}&\sup_{\lambda\in\Lambda_n,1\leqslant i\leqslant n}|\lambda' g_{in}(\theta_0)|\leqslant S_nn^{-1/2}c_n^{-1/2}\max_{1\leqslant i\leqslant n}\|g_{in}(\theta_0)\|\\ &=S_nn^{-1/2}c_n^{-1/2}n^{1/2}S_n^{-1}c_n=c_n^{1/2}=o_p(1),\end{aligned} \tag{A.19}$$

which also immediately implies the second part. □

Lemma 4. *Suppose* $\max_{1\leqslant i\leqslant n}\|g_i(\theta_0)\| = o_p(S_n^{-1}n^{1/2})$, $\lambda_{\min}(\widehat{\Delta}(\theta_0))\geqslant\varepsilon$ *w.p.a.*1 *for some* $\varepsilon>0$, $\widehat{g}_n(\theta_0) = O_p(n^{-1/2})$ *and Assumption* ρ *holds.*

Then $\lambda(\theta_0)\in\widehat{\Lambda}_n(\theta_0)$ *satisfying* $\widehat{P}_\rho(\theta_0,\lambda(\theta_0)) = \sup_{\lambda\in\widehat{\Lambda}_n(\theta_0)}\widehat{P}_\rho(\theta_0,\lambda)$ *exists w.p.a.*1, $\lambda(\theta_0) = O_p(S_n n^{-1/2})$ *and* $\sup_{\lambda\in\widehat{\Lambda}_n(\theta_0)}\widehat{P}_\rho(\theta_0,\lambda) = O_p(S_n n^{-1})$.

Proof. W.l.o.g. $c_n\neq 0$ and thus Λ_n can be assumed compact. Let $\lambda_{\theta_0}\in\Lambda_n$ be such that $\widehat{P}_\rho(\theta_0,\lambda_{\theta_0}) = \max_{\lambda\in\Lambda_n}\widehat{P}_\rho(\theta_0,\lambda)$. Such a $\lambda_{\theta_0}\in\Lambda_n$ exists w.p.a.1 because a continuous function takes on its maximum on a compact set and by Lemma 3 and Assumption ρ, $\widehat{P}_\rho(\theta_0,\lambda)$ (as a function in λ for fixed θ_0) is C^2 on some open neighborhood of Λ_n w.p.a.1. It is now shown that actually $\widehat{P}_\rho(\theta_0,\lambda_{\theta_0}) = \sup_{\lambda\in\widehat{\Lambda}_n(\theta_0)}\widehat{P}_\rho(\theta_0,\lambda)$ w.p.a.1 which then proves the first part of the lemma. By a second-order Taylor expansion around $\lambda = 0$, there is a $\lambda^*_{\theta_0}$ on the line segment joining 0 and λ_{θ_0} such that for some positive constants C_1 and C_2

$$\begin{aligned}0 = S_n\widehat{P}_\rho(\theta_0,0)&\leqslant S_n\widehat{P}_\rho(\theta_0,\lambda_{\theta_0})\\ &= -2S_n\lambda'_{\theta_0}\widehat{g}_n(\theta_0) + \lambda'_{\theta_0}\left[S_n\sum_{i=1}^n \rho_2(\lambda^{*\prime}_{\theta_0}g_{in}(\theta_0))g_{in}(\theta_0)g_{in}(\theta_0)'/n\right]\lambda_{\theta_0} \qquad (A.20)\\ &\leqslant -2S_n\lambda'_{\theta_0}\widehat{g}_n(\theta_0) - C_1\lambda'_{\theta_0}\widehat{\Delta}(\theta_0)\lambda_{\theta_0}\leqslant 2S_n\|\lambda_{\theta_0}\|\ \|\widehat{g}_n(\theta_0)\| - C_2\|\lambda_{\theta_0}\|^2 \qquad (A.21)\end{aligned}$$

w.p.a.1, where the second inequality follows as $\max_{1\leqslant i\leqslant n}\rho_2(\lambda^{*\prime}_{\theta_0}g_{in}(\theta_0)) < -\frac{1}{2}$ w.p.a.1 from Lemma 3, continuity of $\rho_2(\cdot)$ at zero, and $\rho_2 = -1$. The last inequality follows from $\lambda_{\min}(\widehat{\Delta}(\theta_0))\geqslant\varepsilon>0$ w.p.a.1. Now, (A.21) implies that $(C_2/2)\|\lambda_{\theta_0}\|\leqslant S_n\|\widehat{g}_n(\theta_0)\|$ w.p.a.1, the latter being $O_p(S_n n^{-1/2})$ by assumption. It follows that $\lambda_{\theta_0}\in\mathrm{int}(\Lambda_n)$ w.p.a.1. To prove this, let $\varepsilon>0$. Because $\lambda_{\theta_0} = O_p(S_n n^{-1/2})$ and $c_n = o_p(1)$, there exist $M_\varepsilon<\infty$ and $n_\varepsilon\in\mathbb{N}$ such that $\Pr(\|S_n^{-1}n^{1/2}\lambda_{\theta_0}\|\leqslant M_\varepsilon) > 1-\varepsilon/2$ and $\Pr(c_n^{-1/2} > M_\varepsilon) > 1-\varepsilon/2$ for all $n\geqslant n_\varepsilon$. Then $\Pr(\lambda_{\theta_0}\in\mathrm{int}(\Lambda_n)) = \Pr(\|S_n^{-1}n^{1/2}\lambda_{\theta_0}\| < c_n^{-1/2})\geqslant \Pr((\|S_n^{-1}n^{1/2}\lambda_{\theta_0}\|\leqslant M_\varepsilon)\wedge(c_n^{-1/2} > M_\varepsilon)) > 1-\varepsilon$ for $n\geqslant n_\varepsilon$.

Hence, the FOC for an interior maximum $(\partial\widehat{P}_\rho/\partial\lambda)(\theta_0,\lambda) = 0$ hold at $\lambda = \lambda_{\theta_0}$ w.p.a.1. By Lemma 3, $\lambda_{\theta_0}\in\widehat{\Lambda}_n(\theta_0)$ w.p.a.1 and thus by concavity of $\widehat{P}_\rho(\theta_0,\lambda)$ (as a function in λ for fixed θ_0) and convexity of $\widehat{\Lambda}_n(\theta_0)$ it follows that $\widehat{P}_\rho(\theta_0,\lambda_{\theta_0}) = \sup_{\lambda\in\widehat{\Lambda}_n(\theta_0)}\widehat{P}_\rho(\theta_0,\lambda)$ w.p.a.1 which implies the first part of the lemma. From above $\lambda_{\theta_0} = O_p(S_n n^{-1/2})$. Thus the second and by (A.21)the third parts of the lemma follow. □

Proof of Theorem 1. (i) Lemma 4 implies that the FOC

$$n^{-1}\sum_{i=1}^n \rho_1(\lambda' g_{in}(\theta))g_{in}(\theta) = 0 \qquad (A.22)$$

have to hold at $(\theta_0, \lambda_0{:=}\lambda(\theta_0))$ w.p.a.1. Expanding the FOC in λ around 0, there exists a mean value $\widetilde{\lambda}$ between 0 and λ_0 (that may be different for each row) such that

$$0 = -\widehat{g}_n(\theta_0) + \left[S_n\sum_{i=1}^n \rho_2(\widetilde{\lambda}' g_{in}(\theta_0))g_{in}(\theta_0)g_{in}(\theta_0)'/n\right]S_n^{-1}\lambda_0 = -\widehat{g}_n(\theta_0) - \widehat{\Delta}_{\widetilde{\lambda}}S_n^{-1}\lambda_0, \qquad (A.23)$$

where the matrix $\widehat{\Delta}_{\widetilde{\lambda}}$ has been implicitly defined. Because $\lambda_0 = O_p(S_n n^{-1/2})$, Lemma 3 and Assumption ρ imply that $\max_{1\leqslant i\leqslant n}|\rho_2(\widetilde{\lambda}' g_{in}(\theta_0)) + 1|\to_p 0$. By Assumption M_{θ_0}(ii) and Lemma 2 it follows that $\widehat{\Delta}_{\widetilde{\lambda}}\to_p 2\Delta(\theta_0)>0$ and thus $\widehat{\Delta}_{\widetilde{\lambda}}$ is invertible w.p.a.1 and $(\widehat{\Delta}_{\widetilde{\lambda}})^{-1}\to_p\Delta(\theta_0)^{-1}/2$. Therefore,

$$S_n^{-1}\lambda_0 = -(\widehat{\Delta}_{\widetilde{\lambda}})^{-1}\widehat{g}_n(\theta_0) \qquad (A.24)$$

w.p.a.1. Inserting this into a second-order Taylor expansion for $\widehat{P}(\theta,\lambda)$ (with mean value λ^* as in (A.21) above) it follows that w.p.a.1

$$S_n^{-1}n\widehat{P}_\rho(\theta_0,\lambda_0) = 2n\widehat{g}_n(\theta_0)'\widehat{\Delta}_{\widetilde{\lambda}}^{-1}\widehat{g}_n(\theta_0) - n\widehat{g}_n(\theta_0)'\widehat{\Delta}_{\widetilde{\lambda}}^{-1}\widehat{\Delta}_{\lambda^*}\widehat{\Delta}_{\widetilde{\lambda}}^{-1}\widehat{g}_n(\theta_0). \qquad (A.25)$$

By Lemma 1 and M_{θ_0}(iii) $n^{1/2}\widehat{g}_n(\theta_0) = 2n^{1/2}\widehat{g}(\theta_0) + \mathrm{o_p}(1) \rightarrow_{\mathrm{d}} 2\mathrm{N}(0, \Delta(\theta_0))$ and therefore $S_n^{-1} n\widehat{P}_\rho(\theta_0, \lambda_0)/2 \rightarrow_{\mathrm{d}} \chi^2(k)$.

(i)$'$ Note that Assumption M(i)–(iii) in GS (p. 673), for $\Theta = \{\theta_0\}$ is implied by Assumption M_{θ_0}(i)–(iii) above. The result then follows from (2.19) and the proof of Theorem 3 in GS.

(ii) Define $D := D_\rho(\theta_0)\Lambda$ where the $p \times p$ diagonal matrix $\Lambda := \mathrm{diag}(n^{1/2}, \ldots, n^{1/2}, 1, \ldots, 1)$ has first p_A diagonal elements equal to $n^{1/2}$ and the remainder equal to unity. Then (in the remainder of the proof the argument θ_0 is left out for notational simplicity) it follows that

$$LM_\rho = n\widehat{g}_n' \widehat{\Delta}^{-1} D(D'\widehat{\Delta}^{-1} D)^{-1} D'\widehat{\Delta}^{-1}\widehat{g}_n/2. \tag{A.26}$$

It follows from (A.24) and $n^{1/2}\widehat{g}_n = \mathrm{O_p}(1)$ that

$$S_n^{-1} n^{1/2}\lambda_0 = -\Delta^{-1} n^{1/2}\widehat{g}_n/2 + \mathrm{o_p}(1) \tag{A.27}$$

and therefore the statement of the theorem involving S_ρ follows immediately from the one for LM_ρ. Therefore, only the statistic LM_ρ is dealt with using its representation in Eq. (A.26).

First, it is shown that the matrix D is asymptotically independent of $n^{1/2}\widehat{g}_n$. By a mean-value expansion about 0 it follows that $\rho_1(\lambda_0' g_{in}) = -1 + \rho_2(\xi_i) g_{in}'\lambda_0$ for a mean value ξ_i between 0 and $\lambda_0' g_{in}$ and thus by (2.14), (A.27), and the definition of Λ it follows that (modulo $\mathrm{o_p}(1)$ terms)

$$\begin{aligned} D &= -n^{-1}\sum_{i=1}^n (n^{1/2} G_{inA}, G_{inB}) - S_n n^{-3/2}\sum_{i=1}^n [\rho_2(\xi_i)(n^{1/2}G_{inA}, G_{inB}) g_{in}'\Delta^{-1} n^{1/2}\widehat{g}_n]/2 \\ &= -\left(n^{-1/2}\sum_{i=1}^n G_{inA} - S_n n^{-1}\sum_{i=1}^n G_{inA} g_{in}'\Delta^{-1} n^{1/2}\widehat{g}_n/2, 2M_2(\beta_0)\right), \end{aligned} \tag{A.28}$$

where for the last equality we use (A.5) and Assumptions M_{θ_0}(v)–(vi). By Assumption M_{θ_0}(v) and Eq. (A.13) it follows that $\widehat{\Delta}_A = S_n n^{-1}\sum_{i=1}^n vec(G_{inA}) g_{in}'/2 \rightarrow_{\mathrm{p}} \Delta_A$ and thus

$$vec(D, n^{1/2}\widehat{g}_n) = w_1 + Mv + \mathrm{o_p}(1), \quad \text{where} \tag{A.29}$$

$$w_1 := vec(0, -2M_2(\beta_0), 0) \in \mathbb{R}^{kp_A + kp_B + k} \quad \text{and}$$
$$M := \begin{pmatrix} -I_{kp_A} & \Delta_A\Delta^{-1} \\ 0 & 0 \\ 0 & I_k \end{pmatrix}, \quad v := n^{-1/2}\sum_{i=1}^n \begin{pmatrix} vec\, G_{inA} \\ g_{in} \end{pmatrix}; \tag{A.30}$$

M and v have dimensions $(kp_A + kp_B + k) \times (kp_A + k)$ and $(kp_A + k) \times 1$, respectively. By Assumption ID, M_{θ_0}(vii), Lemma 1, and (A.5) it follows that $v \rightarrow_{\mathrm{d}} 2\mathrm{N}(w_2, V)$, where

$$w_2 := ((vec M_{1A})', 0)' \tag{A.31}$$

and M_{1A} are the first p_A columns of M_1. Therefore

$$vec(D, n^{1/2}\widehat{g}_n) \rightarrow_{\mathrm{d}} \mathrm{N}\left(w_1 + 2Mw_2, 4\begin{pmatrix} \Psi & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \Delta \end{pmatrix}\right), \tag{A.32}$$

where $\Psi := \Delta_{AA} - \Delta_A\Delta^{-1}\Delta_A'$ has full column rank. Eq. (A.32) proves that D and $n^{1/2}\widehat{g}_n$ are asymptotically independent.

The asymptotic distribution of LM_ρ is derived next. Denote by $\overline{D}$ and $\overline{g}$ the limiting normal random matrices corresponding to D and $n^{1/2}\widehat{g}_n$, respectively, see (A.32). Below it is shown that the function $h : \mathbb{R}^{k\times p} \rightarrow \mathbb{R}^{p\times k}$ defined by $h(d) := (d'\Delta^{-1}d)^{-1/2}d'$ for $d \in \mathbb{R}^{k\times p}$ is continuous on a set $C \subset \mathbb{R}^{k\times p}$ with

$\Pr(\overline{D} \in C) = 1$. By the continuous mapping theorem and $\mathrm{M}_{\theta_0}(\mathrm{v})$ it follows that

$$2^{-1/2}(D'\widehat{\Delta}^{-1}D)^{-1/2}D'\widehat{\Delta}^{-1}n^{1/2}\widehat{g}_n \rightarrow_{\mathrm{d}} (\overline{D}'\Delta^{-1}\overline{D})^{-1/2}\overline{D}'\Delta^{-1}\overline{g}/2. \tag{A.33}$$

By the independence of $\overline{D}$ and $\overline{g}$, the latter random variable is distributed as ζ, where $\zeta \sim \mathrm{N}(0, I_p)$.

Finally, the continuity claim for h is dealt with. Note that h is continuous at each $d \in \mathbb{R}^{k\times p}$ that has full column rank. It is therefore sufficient to show that $\overline{D}$ has full column rank a.s. From (A.32) it follows that the last p_B columns of $\overline{D}$ equal $-2M_2(\beta_0)$ which has full column rank by assumption. Define $O := \{o \in \mathbb{R}^{kp_A} : \exists \widetilde{o} \in \mathbb{R}^{k\times p_A}$, s.t. $o = vec(\widetilde{o})$ and the $k \times p$ matrix $(\widetilde{o}, -2M_2(\beta_0))$ has linearly dependent columns$\}$. Clearly, O is closed and therefore Lebesgue measurable. Furthermore, O has empty interior and thus has Lebesgue measure 0. For the first p_A columns of $\overline{D}$, $\overline{D}_{p_A}$ say, it has been shown that $vec\,\overline{D}_{p_A}$ is normally distributed with full rank covariance matrix Ψ. This implies that for any measurable set $O^+ \subset \mathbb{R}^{kp_A}$ with Lebesgue measure 0,$\Pr(vec(\overline{D}_{p_A}) \in O^+) = 0$, in particular, for $O^+ = O$. This proves the continuity claim for h.

(ii)′ Note that under (2.19) the analogue to (A.8) in GS is $n^{1/2}\lambda_0 = -\Omega^{-1}n^{1/2}\widehat{g} + \mathrm{o_p}(1)$. It follows that S^*_ρ and LM^*_ρ have the same asymptotic distribution and it is thus enough to prove the result for LM^*_ρ. As in the proof of Theorem 4 in GS (line 12↑, p. 706) we have a formula

$$vec(D^*, n^{1/2}\widehat{g}) = w_1^* + M^*v + \mathrm{o_p}(1) \tag{A.34}$$

with D^* defined in GS (line 10↑, p. 681) $w_1^* = vec(0, -M_2(\beta_0), 0)$, and $v^* = n^{-1/2}\sum_{i=1}^n ((vec\,G_{iA})', g_i')'$ but, because of (2.19) and (A.10), we have—in contrast to GS—that

$$M^* := \begin{pmatrix} -I_{kp_A} & \Omega_A\Omega^{-1} \\ 0 & 0 \\ 0 & I_k \end{pmatrix} \in \mathbb{R}^{(kp_A+kp_B+k)+(kp_A+k)}. \tag{A.35}$$

By Assumption $\mathrm{M}_{\theta_0}(\mathrm{vii})$ we have $v^* \rightarrow_{\mathrm{d}} \mathrm{N}(w_2, V)$, with w_2 defined in (A.31). Thus $vec(D^*, n^{1/2}\widehat{g}) \rightarrow_{\mathrm{d}} \zeta$, where ζ is a random variable distributed as $\mathrm{N}(w_1^* + M^*w_2, M^*VM^{*'})$. Because in general $-\Delta_A + \Omega_A\Omega^{-1}\Delta \neq 0$, it follows that D^* and $n^{1/2}\widehat{g}$ are typically not asymptotically independent. Therefore, in general, $n^{1/2}\widehat{g}$ is no longer asymptotically $\mathrm{N}(0, \Delta)$ conditional on D^*, and consequently LM^*_ρ is not asymptotically χ^2. More specifically, let $\zeta_1 \in \mathbb{R}^{k\times p}$ and $\zeta_2 \in \mathbb{R}^k$ be random matrices such that $\zeta = ((vec\,\zeta_1)', \zeta_2')'$. It then follows that

$$LM^*_\rho \rightarrow_{\mathrm{d}} \widetilde{\xi} := \zeta_2'\Omega^{-1}\zeta_1(\zeta_1'\Omega^{-1}\zeta_1)^{-1}\zeta_1'\Omega^{-1}\zeta_2. \qquad \square \tag{A.36}$$

Proof of Theorem 2. (i) We first show that $\widehat{\beta} \rightarrow_{\mathrm{p}} \beta_0$. Note that Assumptions $\mathrm{M}_{\alpha_0}(\mathrm{i})$, (ii), and (iv) do not assume consistency of $\widehat{\beta}$. A proof as for Lemma 3 using the first portion of $\mathrm{M}_{\alpha_0}(\mathrm{i})$ shows that

$$\sup_{\beta\in B, \lambda\in\Lambda_n, 1\leqslant i\leqslant n} |\lambda' g_{in}(\theta_\beta)| \rightarrow_{\mathrm{p}} 0, \tag{A.37}$$

where the definition of c_n is changed to $c_n = S_n n^{-1/2}\max_{1\leqslant i\leqslant n}\sup_{\beta\in B}\|g_{in}(\theta_\beta)\|$. By $\sup_{\lambda\in\widehat{\Lambda}_n(\theta_0)}\widehat{P}_\rho(\theta_0, \lambda) = \mathrm{O_p}(S_n n^{-1})$ (which holds by Lemma 4) a variant of Lemma 9 in GS (defining $\underline{\lambda} := -S_n n^{-1/2}\widehat{g}_n(\widehat{\theta}_0)/\|\widehat{g}_n(\widehat{\theta}_0)\|$ in their proof, using $\sup_{\lambda\in\Lambda_n, 1\leqslant i\leqslant n}|\rho_2(\lambda' g_{in}(\widehat{\theta}_0)) + 1| \rightarrow_{\mathrm{p}} 0$ which holds by (A.37), and using $\lambda_{\max}(\widehat{\Delta}(\widehat{\theta}_0))$ bounded w.p.a.1 which holds by $\mathrm{M}_{\alpha_0}(\mathrm{ii})$) yields $\widehat{g}_n(\widehat{\theta}_0) = \mathrm{O_p}(n^{-1/2})$. Using Eq. (A.12) in the proof of Lemma 1 (with g_i and g_{in} replaced by $g_i(\widehat{\theta}_0)$ and $g_{in}(\widehat{\theta}_0)$, respectively) and $\mathrm{M}_{\alpha_0}(\mathrm{i})$, $\widehat{g}_n(\widehat{\theta}_0) = \mathrm{O_p}(n^{-1/2})$ implies that $\widehat{g}(\widehat{\theta}_0) = \mathrm{o_p}(1)$. By the first part of $\mathrm{M}_{\alpha_0}(\mathrm{iv})$ and ID_{α_0} we have $\mathrm{o_p}(1) = \widehat{g}(\widehat{\theta}_0) = m_2(\alpha_{02}, \widehat{\beta}) + \mathrm{o_p}(1)$ which by ID_{α_0} implies $\widehat{\beta} \rightarrow_{\mathrm{p}} \beta_0$.

A variant of Lemma 4 using $\widehat{g}_n(\widehat{\theta}_0) = \mathrm{O_p}(n^{-1/2})$ shows that $\widehat{\lambda} := \lambda(\widehat{\theta}_0)$ exists and thus an FOC of $\widehat{P}_\rho(\widehat{\theta}_0, \lambda)$ w.r.t. λ holds w.p.a.1. An analysis as in the proof of Theorem 5 of GS using Eq. (A.20) and an analogue of Eq. (A.27) then yields

$$GELR^{\mathrm{sub}}_\rho(\alpha_0) = n^{1/2}\widehat{g}_n(\widehat{\theta}_0)'\Delta(\theta_0)^{-1}n^{1/2}\widehat{g}_n(\widehat{\theta}_0)/4 + \mathrm{o_p}(1). \tag{A.38}$$

Using the implicit function and envelope theorems the FOC in θ, $0=n^{-1}\sum_{i=1}^{n}\rho_1(\widehat{\lambda}' g_{in}(\widehat{\theta}_0))\,(\partial g_{in}/\partial\beta)'(\widehat{\theta}_0)S_n^{-1}\widehat{\lambda}$, has to hold. Combining this with a mean-value expansion of (A.22) in (β,λ) about $(\beta_0,0)$ we get

$$\begin{pmatrix} 0 \\ -\widehat{g}_n(\theta_0) \end{pmatrix} + M\begin{pmatrix} \widehat{\beta}-\beta_0 \\ S_n^{-1}\widehat{\lambda} \end{pmatrix} = 0, \tag{A.39}$$

where

$$M{:=}n^{-1}\sum_{i=1}^{n}\begin{pmatrix} 0 & \rho_1(\widehat{\lambda}' g_{in}(\widehat{\theta}_0))(\partial g_{in}/\partial\beta)'(\widehat{\theta}_0) \\ \rho_1(\overline{\lambda}' g_{in}(\widehat{\theta}_0))(\partial g_{in}/\partial\beta)(\theta_{\overline{\beta}}) & S_n\rho_2(\overline{\lambda}' g_{in}(\widehat{\theta}_0))g_{in}(\theta_{\overline{\beta}})g_{in}(\widehat{\theta}_0)' \end{pmatrix} \tag{A.40}$$

and $(\overline{\beta}',\overline{\lambda}')$ are mean values on the line segment joining $(\widehat{\beta},\widehat{\lambda})$ and $(\beta_0',0')$. Note that by the last two conditions in M_{α_0}(iii) and by an analysis as in (A.12) in the proof of Lemma 1, we have $(\widehat{G}_{nB}-2\widehat{G}_B)(\theta_{\widetilde{\beta}})=\mathrm{o_p}(1)$ for any argument $\theta_{\widetilde{\beta}}$ as in M_{α_0}. Again by M_{α_0}(iii), we have $\widehat{G}_B(\theta_{\widetilde{\beta}})\to_{\mathrm{p}}M_{2\beta}(\alpha_{02},\beta_0)$, where $\mathrm{M}_{2\beta}(\cdot){:=}(\partial m_2/\partial\beta)(\cdot)\in\mathbb{R}^{k\times p_B}$.Therefore by M_{α_0}(ii), $M\to_{\mathrm{p}}\overline{M}$, where (writing $\mathrm{M}_{2\beta}$ for $\mathrm{M}_{2\beta}(\alpha_{02},\beta_0)$ and Δ for $\Delta(\theta_0)$)

$$\overline{M}{:=}-2\begin{pmatrix} 0 & M_{2\beta}' \\ M_{2\beta} & \Delta \end{pmatrix},\quad \overline{M}^{-1}=-2^{-1}\begin{pmatrix} -\Sigma & H \\ H' & P \end{pmatrix},$$

$$\Sigma{:=}(M_{2\beta}'\Delta^{-1}M_{2\beta})^{-1},\quad H{:=}\Sigma M_{2\beta}'\Delta^{-1}\quad\text{and}\quad P{:=}\Delta^{-1}-\Delta^{-1}M_{2\beta}\Sigma M_{2\beta}'\Delta^{-1}. \tag{A.41}$$

By (A.39) w.p.a.1

$$n^{1/2}((\widehat{\beta}-\beta_0)',S_n^{-1}\widehat{\lambda}')'=M^{-1}(0',n^{1/2}\widehat{g}_n(\theta_0)')'=2M^{-1}(0',n^{1/2}\widehat{g}(\theta_0)')'+\mathrm{o_p}(1), \tag{A.42}$$

where the second equality holds by Lemma 1 using M_{α_0}(i) and (iv). An expansion of $\widehat{g}(\widehat{\theta}_0)$ in β around β_0 and the above lead to (up to $\mathrm{o_p}(1)$ terms)

$$n^{1/2}\widehat{g}_n(\widehat{\theta}_0)=n^{1/2}2\widehat{g}(\widehat{\theta}_0)=n^{1/2}2[\widehat{g}(\theta_0)+\widehat{G}_B(\overline{\theta})(\widehat{\beta}-\beta_0)]=2(I_k-M_{2\beta}H)n^{1/2}\widehat{g}(\theta_0) \tag{A.43}$$

for some appropriate mean value $\overline{\theta}$, where the first equality can be established by an analogous expansion for $n^{1/2}\widehat{g}_n(\widehat{\theta}_0)$, Lemma 1, and $n^{1/2}(\widehat{\beta}-\beta_0)=\mathrm{O_p}(1)$. Note that $\mathrm{M}_{M_{2\beta}}(\Delta)=I_k-M_{2\beta}H$ and $\Delta^{-1/2}M_{M_{2\beta}}(\Delta)\Delta^{1/2}=M_{\Delta^{-1/2}M_{2\beta}}$. Then, by (A.38) and M_{α_0}(iv), $GELR_\rho^{\mathrm{sub}}(\alpha_0)\to_{\mathrm{d}}\xi' M_{\Delta^{-1/2}M_{2\beta}}\xi$ for $\xi\sim\mathrm{N}(0,I_k)$ and since $\Delta^{-1/2}M_{2\beta}$ is of rank p_B we obtain $GELR_\rho^{\mathrm{sub}}(\alpha_0)\to_{\mathrm{d}}\chi^2(k-p_B)$ as claimed.

(ii) By a modification of (A.27), the result for $LM_\rho^{\mathrm{sub}}(\alpha_0)$ implies the result for $S_\rho^{\mathrm{sub}}(\alpha_0)$. Renormalize $D{:=}D_\rho(\alpha_0)\Lambda$, where $\Lambda{:=}\mathrm{diag}(n^{1/2},\ldots,n^{1/2},1,\ldots,1)$ has p_{A_1} elements equal to $n^{1/2}$ and p_{A_2} elements equal to 1. The key portion of the proof is to show asymptotic independence of D and $n^{1/2}\widehat{g}_n(\widehat{\theta}_0)$. By a mean-value expansion about θ_0 we have for a mean value $\theta_{\widetilde{\beta}}$ (that may be different for each row) and (A.42)

$$\begin{aligned} n^{1/2}\,vec\,\widehat{G}_{A_1}(\widehat{\theta}_0)&=n^{1/2}\,vec\,\widehat{G}_{A_1}(\theta_0)+(\partial\,vec\,\widehat{G}_{A_1}/\partial\beta)(\theta_{\widetilde{\beta}})n^{1/2}(\widehat{\beta}-\beta_0)\\ &=n^{1/2}\,vec\,\widehat{G}_{A_1}(\theta_0)-(\partial\,vec\,\widehat{G}_{A_1}/\partial\beta)(\theta_{\widetilde{\beta}})Hn^{1/2}\widehat{g}(\theta_0)+\mathrm{o_p}(1)\\ &=n^{1/2}\,vec\,\widehat{G}_{A_1}(\theta_0)+\mathrm{o_p}(1) \end{aligned} \tag{A.44}$$

by Assumption M_{α_0}(v). By M_{α_0}(vi) we thus have $vec\,\widehat{G}_{A_1}(\widehat{\theta}_0)=\mathrm{O_p}(n^{-1/2})$. Then, by an analysis as in Lemma 1 and the first part of M_{α_0}(vi) it follows that

$$n^{1/2}\,vec\,\widehat{G}_{nA_1}(\widehat{\theta}_0)=n^{1/2}2\,vec\,\widehat{G}_{A_1}(\theta_0)+\mathrm{o_p}(1). \tag{A.45}$$

By M_{α_0}(vii), (A.43), and (A.45) it then follows that

$$vec(D,n^{1/2}\widehat{g}_n(\widehat{\theta}_0))=2m+2Mv+\mathrm{o_p}(1), \tag{A.46}$$

where $M \in \mathbb{R}^{(kp_{A_1}+kp_{A_2}+k)\times(kp_{A_1}+k)}$ and

$$M := \begin{pmatrix} -I_{kp_{A_1}} & \Delta_{A_1}\Delta^{-1} \\ 0 & 0 \\ 0 & I_k \end{pmatrix} \begin{pmatrix} I_{kp_{A_1}} & 0 \\ 0 & I_k - M_{2\beta}H \end{pmatrix},$$

$$v := n^{-1/2}\sum_{i=1}^{n}\begin{pmatrix} vec\, G_{iA_1}(\theta_0) \\ g_i(\theta_0) \end{pmatrix}, \quad m := vec(0, -(\partial m_2/\partial\alpha_2), 0), \tag{A.47}$$

where the arguments (α_{02}, β_0) in $M_{2\beta}$ and $(\partial m_2/\partial\alpha_2)$ and θ_0 in Δ_{A_1} and Δ are omitted. Note here that the last two conditions in M_{α_0}(vii) and analysis as in Lemma 1 imply $\widehat{G}_{nA_1}(\widehat{\theta}_0) - 2\widehat{G}_{A_1}(\widehat{\theta}_0) = o_p(1)$. By M_{α_0}(vi), v is asymptotically normal with full rank covariance matrix V^α and thus the asymptotic covariance matrix of $vec(D, n^{1/2}\widehat{g}_n(\widehat{\theta}_0))$ is given by $4MV^\alpha M'$. For independence of D and $n^{1/2}\widehat{g}_n(\widehat{\theta}_0)$ the upper right $k(p_{A_1} + p_{A_2}) \times k$ submatrix of $MV^\alpha M'$ must be 0. This is clear for the $kp_{A_2} \times k$-dimensional submatrix and we only have to show that the $kp_{A_1} \times k$ upper right submatrix

$$[-\Delta_{A_1} + \Delta_{A_1}\Delta^{-1}(I_k - M_{2\beta}H)\Delta](I_k - M_{2\beta}H)' \tag{A.48}$$

is 0. Using $I_k - M_{2\beta}H = M_{M_{2\beta}}(\Delta)$, the matrix in (A.48) equals $-\Delta_{A_1}\Delta^{-1}P_{M_{2\beta}}(\Delta)M_{M_{2\beta}}(\Delta)\Delta$ which is clearly 0. This proves the independence claim. Denote by $\overline{D}$ and $\overline{g}$ the limiting normal distributions of D and $n^{1/2}\widehat{g}_n(\widehat{\theta}_0)$, implied by (A.46). Set $M^* = \Delta^{-1}M_{M_{2\beta}}(\Delta)$ and note that $2\widehat{M}(\alpha_0)\to_p M^*$ for the matrix in (2.26). The function $h: \mathbb{R}^{k\times p_A} \to \mathbb{R}^{p_A\times k}$ defined by $h(d) := (d'M^*d)^{-1/2}d'$ for $d \in \mathbb{R}^{k\times p_A}$ is continuous on a set $C \subset \mathbb{R}^{k\times p_A}$ with $\Pr(\overline{D} \in C) = 1$ (which is proved along the same lines as in Theorem 1). By the continuous mapping theorem and (A.43)

$$(D'M^*D)^{-1/2}D'\Delta^{-1}n^{1/2}\widehat{g}_n(\widehat{\theta}_0)\to_d(\overline{D}'M^*\overline{D})^{-1/2}\overline{D}'\Delta^{-1}\overline{g}\sim 2\mathrm{N}(0, I_{p_A}). \tag{A.49}$$

Because $\widehat{\Delta}(\widehat{\theta}_0)\to_p 2\Delta$ the claim follows. $\square$

Proof of Theorem 3. Let $\mu_0 := \mu_\rho(\theta_0)$. Inserting this into a second-order Taylor expansion for $\widehat{P}_\rho(\theta, \mu)$ around $\mu = 0$ with mean value $\widetilde{\mu}$, cf. Eq. (A.21) above,

$$\begin{aligned} S_n\widehat{P}_\rho(\theta_0, \mu_0) &= -2S_n\mu_0'\widehat{g}_n(\theta_0) + \mu_0'\left[S_n\sum_{i=1}^{n}\rho_2(\widetilde{\mu}'g_{in}(\theta_0))g_{in}(\theta_0)g_{in}(\theta_0)'/n\right]\mu_0 \\ &= -2S_n\mu_0'\widehat{g}_n(\theta_0) + \mu_0'\widehat{\Delta}_{\widetilde{\mu}}\mu_0, \end{aligned} \tag{A.50}$$

where $\widehat{\Delta}_{\widetilde{\mu}}$ has been implicitly defined. As in the proof of Theorem 1(ii) define $D := D_\rho(\theta_0)\Lambda$. Hence, we may write $\mu_0 = -S_n\widehat{\Delta}(\theta_0)^{-1}D(D'\widehat{\Delta}(\theta_0)^{-1}D)^{-1}D'\widehat{\Delta}(\theta_0)^{-1}\widehat{g}_n(\theta_0)$. From Assumption M_{θ_0}(ii) and Lemma 2, both $\lambda_{\min}(\widehat{\Delta}(\theta_0))$ and $\lambda_{\min}(\widehat{\Delta}(\theta_0)^{-1}) \geqslant \varepsilon > 0$ w.p.a.1. Therefore, as the expression in (A.33) and D are $O_p(1)$, it follows that $\mu_0 = O_p(S_n n^{-1/2})$. By an analogous argument to that in the proof of Lemma 4, $\mu_0 \in \mathrm{int}(\Lambda_n)$ w.p.a.1. Therefore, Lemma 3 and Assumption ρ imply that $\max_{1\leqslant i\leqslant n}|\rho_2(\widetilde{\mu}'g_{in}(\theta_0)) + 1|\to_p 0$ and, thus from the last part of Assumption M_{θ_0}(ii), $\widehat{\Delta}_{\widetilde{\mu}}\to_p -2\Delta(\theta_0)$. Thus, substituting for μ_0,

$$\begin{aligned} S_n^{-1}n\widehat{P}_\rho(\theta_0, \mu_0) &= n\widehat{g}_n(\theta_0)'\widehat{\Delta}(\theta_0)^{-1}D(D'\widehat{\Delta}(\theta_0)^{-1}D)^{-1}D'\widehat{\Delta}(\theta_0)^{-1}\widehat{g}_n(\theta_0) + o_p(1) \\ &= 2LM_\rho(\theta_0) + o_p(1)\to_d 2\chi^2(p), \end{aligned} \tag{A.51}$$

from the proof of Theorem 1(ii) as $\widehat{\Delta}\to_p 2\Delta(\theta_0)$ and by Lemma 1 and M_{θ_0}(iii) $n^{1/2}\widehat{g}_n(\theta_0) = 2n^{1/2}\widehat{g}(\theta_0) + o_p(1)\to_d 2\mathrm{N}(0, \Delta(\theta_0))$. The result for $S_n^{-1}n\widehat{P}_\rho(\theta_0, \widetilde{\mu}(\theta_0))/2$ then also follows immediately as $\lambda(\theta_0) = -S_n\widehat{\Delta}(\theta_0)^{-1}\widehat{g}_n(\theta_0) + o_p(S_n n^{-1/2})$. $\square$

References

Anderson, T.W., Rubin, H., 1949. Estimators of the parameters of a single equation in a complete set of stochastic equations. The Annals of Mathematical Statistics 21, 570–582.

Andrews, D.W.K., 1991. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. Econometrica 59 (3), 817–858.

Andrews, D.W.K., Guggenberger, P., 2005a. The limit of finite-sample size and a problem with subsampling. Cowles Foundation Discussion Paper 1605.

Andrews, D.W.K., Guggenberger, P., 2005b. Hybrid and size-corrected subsample methods. Cowles Foundation Discussion Paper 1606.

Andrews, D.W.K., Guggenberger, P., 2005c. Applications of subsampling, hybrid, and size-correction methods. Unpublished manuscript, Cowles Foundation, Yale University.

Andrews, D.W.K., Marmer, V., 2004. Exactly distribution-free inference in instrumental variables regression with possibly weak instruments. Journal of Econometrics, forthcoming.

Andrews, D.W.K., Moreira, M., Stock, J.H., 2006. Optimal invariant similar tests for instrumental variables regression. Econometrica 74 (3), 715–752.

Caner, M., 2003. Exponential tilting with weak instruments: estimation and testing. Unpublished manuscript, Department of Economics, Pittsburgh University.

Chao, J.C., Swanson, N.R., 2005. Consistent estimation with a large number of weak instruments. Econometrica 73 (5), 1673–1692.

Dufour, J., 1997. Some impossibility theorems in econometrics with applications to structural and dynamic models. Econometrica 65 (6), 1365–1387.

Dufour, J., 2003. Identification, weak instruments and statistical inference in econometrics. Presidential address to the Canadian Economics Association. Canadian Journal of Economics 36 (4), 767–808.

Dufour, J., Taamouti, M., 2005. Projection-based statistical inference in linear structural models with possibly weak instruments. Econometrica 73 (4), 1351–1365.

Guggenberger, P., 2003. Econometric essays on generalized empirical likelihood, long-memory time series, and volatility. Ph.D. Thesis, Yale University.

Guggenberger, P., Smith, R.J., 2005. Generalized empirical likelihood estimators and tests under partial, weak and strong identification. Econometric Theory 21 (4), 667–709.

Guggenberger, P., Wolf, M., 2004. Subsampling tests of parameter hypotheses and overidentifying restrictions with possible failure of identification. Unpublished manuscript, Department of Economics, UCLA.

Hamilton, J.D., 1994. Time Series Analysis. Princeton University Press, Princeton, NJ.

Hannan, E.J., 1957. The variance of the mean of a stationary process. Journal of the Royal Statistical Society, Series B 19 (2), 282–285.

Hannan, E.J., 1970. Multiple Time Series. Wiley, New York.

Hansen, L.P., 1982. Large sample properties of generalized method of moments estimators. Econometrica 50 (4), 1029–1054.

Hansen, L.P., Heaton, J., Yaron, A., 1996. Finite-sample properties of some alternative GMM estimators. Journal of Business and Economic Statistics 14 (3), 262–280.

Imbens, G., 1997. One-step estimators for over-identified generalized method of moments models. Review of Economic Studies 64, 359–383.

Imbens, G., Spady, R.H., Johnson, P., 1998. Information theoretic approaches to inference in moment condition models. Econometrica 66 (2), 333–357.

Jansson, M., 2002. Consistent covariance matrix estimation for linear processes. Econometric Theory 18, 1449–1459.

Kiefer, N., Vogelsang, T., 2002a. Heteroskedasticity–autocorrelation robust testing using bandwidth equal to sample size. Econometric Theory 18, 1350–1366.

Kiefer, N., Vogelsang, T., 2002b. Heteroskedasticity–autocorrelation robust standard errors using the Bartlett kernel without truncation. Econometrica 70, 2093–2095.

Kiefer, N., Vogelsang, T., 2005. A new asymptotic theory for heteroskedasticity–autocorrelation robust tests. Econometric Theory 21, 1130–1164.

Kiefer, N., Vogelsang, T., Bunzel, H., 2000. Simple robust testing of regression hypothesis. Econometrica 68 (3), 695–714.

Kitamura, Y., 1997. Empirical likelihood methods with weakly dependent processes. Annals of Statistics 25 (5), 2084–2102.

Kitamura, Y., Stutzer, M., 1997. An information-theoretic alternative to generalized method of moments estimation. Econometrica 65 (4), 861–874.

Kleibergen, F., 2002. Pivotal statistics for testing structural parameters in instrumental variables regression. Econometrica 70 (5), 1781–1805.

Kleibergen, F., 2004. Testing subsets of structural parameters in the instrumental variables regression model. Review of Economics and Statistics 86, 418–423.

Kleibergen, F., 2005a. Testing parameters in GMM without assuming that they are identified. Econometrica 73 (4), 1103–1123.

Kleibergen, F., 2005b. Subset statistics in the linear IV regression model. Unpublished manuscript, Department of Economics, Brown University.

Moreira, M.J., 2003. A conditional likelihood ratio test for structural models. Econometrica 71 (4), 1027–1048.

Moreira, M.J., Porter, J., Suarez, G.A., 2005a. Bootstrap and higher-order expansion validity when instruments may be weak. Unpublished manuscript, Department of Economics, Harvard University.

Moreira, M.J., Porter, J., Suarez, G.A., 2005b. Bootstrap validity for the score test when instruments may be weak. Unpublished manuscript, Department of Economics, Harvard University.

Newey, W.K., Smith, R.J., 2004. Higher order properties of GMM and generalized empirical likelihood estimators. Econometrica 72 (1), 219–255.

Newey, W.K., West, K.D., 1994. Automatic lag selection in covariance matrix estimation. Review of Economic Studies 61, 631–653.

Otsu, T., 2006. Generalized empirical likelihood inference for nonlinear and time series models under weak identification. Econometric Theory 22 (3), 513–527.

Parzen, E., 1956. On consistent estimates of the spectral density of a stationary time series. Proceedings of the National Academy of Sciences 42, 154–157.

Parzen, E., 1957. On consistent estimates of the spectrum of a stationary time series. The Annals of Mathematical Statistics 28, 329–348.

Qin, J., Lawless, J., 1994. Empirical likelihood and general estimating equations. Annals of Statistics 22 (1), 300–325.

Smith, R.J., 1997. Alternative semi-parametric likelihood approaches to generalized method of moments estimation. Economic Journal 107 (441), 503–519.

Smith, R.J., 2000. Empirical likelihood estimation and inference. In: Salmon, M., Marriott, P. (Eds.), Applications of Differential Geometry to Econometrics. Cambridge University Press, Cambridge, pp. 119–150.

Smith, R.J., 2001. GEL criteria for moment condition models. Unpublished manuscript, University of Bristol. Revised version: CWP 19/04, cemmap, I.F.S. and U.C.L. ⟨http://www.cemmap.ifs.org.uk/wps/cwp0419.pdf⟩.

Smith, R.J., 2005. Automatic positive semi-definite HAC covariance matrix and GMM estimation. Econometric Theory 21 (1), 158–170.

Staiger, D., Stock, J.H., 1997. Instrumental variables regression with weak instruments. Econometrica 65 (3), 557–586.

Stock, J.H., Wright, J.H., 2000. GMM with weak identification. Econometrica 68 (5), 1055–1096.

Wooldridge, J.M., White, H., 1988. Some invariance principles and central limit theorems for dependent heterogeneous processes. Econometric Theory 4, 210–230.